The study of DNA methylation anomalies

in chronic lymphocytic leukaemia

Noémi BA Roy

St John's College



A thesis submitted for the degree of

Doctor of Philosophy

of

University of Oxford

MRC Molecular Haematology Unit
Weatherall Institute of Molecular Medicine
University of Oxford

Abstract

Noémi BA Roy
St John's College

Doctor of Philosophy
Michaelmas Term 2011

Many haematological malignancies are associated with widespread alterations of the transcriptional and epigenetic programmes. Changes in DNA methylation provide the clearest example of epigenetic changes, but the mechanism(s) underlying such changes is unknown. To investigate this I studied DNA methylation across an ~80kb segment of the genome which is not known to be mutated in haematological malignancies. Methylation was perturbed in 35-100% of samples of DNA from individuals with a wide range of haematological malignancies but not in non-malignant haematological disorders. DNA methylation was comprehensively assessed by Southern blot analysis, classical bisulphite sequencing and using a newly developed capture bisulphite sequencing protocol. The results were also compared with analysis by MeDIP, an immunoprecipitation-based technique. These analyses provide methylation status at various levels including individual CpG resolution. This showed both gain and loss of methylation at CpG dinucleotides. Of interest, hypomethylation was most frequently seen in intergenic regions corresponding to transcription factor binding sites and areas of increased chromosome accessibility. These observations suggested that hypomethylation of the genome in haematological malignancies could arise from aberrantly expressed DNA binding proteins which, recruited to sequences in regions of open chromatin, would protect the underlying CpG dinucleotides from the methylation machinery. This, in turn, could lead to passive demethylation accumulating with increasing cell divisions. This hypothesis was tested with electrophoretic mobility shift assays using oligonucleotides representing the DNA underlying one such region. This showed that, compared to nuclear extracts from the lymphocytes of normal individuals, those from patients with CLL were enriched for a protein which binds to oligonucleotides containing the underlying sequence. Using a mass spectrometry approach, I identified a variety of proteins that may bind such regions and account for their passive demethylation in haematological malignancies.

Acknowledgements

To Doug Higgs and Richard Gibbons, thank you for giving me the opportunity to carry out this work. I am grateful not only for the experience, but also that you have shared with me the passion that comes from just wanting to understand how things work. Your steadfast rigour has been a pleasure to try to live up to.

To David Garrick, a debt of gratitude for your kindness and patience in casting a close eye over my work, and for your invaluable suggestions, comments and questions.

To Helena Ayyub, for all your help, especially the beautiful gel shifts. To past and present lab members: David Clynes, Marco de Gobbi, Chris Fisher, Deborah Hay, Clare Jelinska, Martin Law, Karen Lower, Magnus Lynch, Mary Muers, Diu Nguyen, Andrew Smith, Douglas Vernimmen, Hsiao Voon for your help and critical assessment of my work over time. To Liz Rose for helping all the balls stay in the air. Special thanks to Jim Hughes and Nicky Gray for all bioinformatics help, and to Monika Kowalczyk for much advice, and whose friendship extended to numerous occasions of emergency childcare! Special thanks to Caroline Scott and Ricarda Gaentzsch for support and "methylation journal clubs" and the opportunity to discuss results and directions to take.

To Anna Schuh, Maite Cabes, Chris Levett and all the participating CLL patients. To Jackie Sloane-Stanley for help with sample preparation and tissue culture.

To Vardhman Rakyan and Skirmantas Kriaucionis for your time, advice and use of your labs, and to Thomas Down and Felix Krueger for bioinformatics support.

To all our supportive friends in Oxford- Mira and Mo, Raffaele and Chantal, Coralie, and to Maria and Lisa, for your invaluable help with the girls.

To Edward **H**ibbert for all the calculating web pages, and for helping me get those dots 'just so'. You see, you did find a way to give me extra hours in the day. Thank you for wanting to listen to *every*thing in **g**ory detail.

To my parents, Françoise and Michel Roy, for not only putting up with all the "why" questions that started so early, but for positively encouraging them. To Anna, Marie, Nic, Gabrielle and Mal for the constant support and cheerful encouragement and believing in me so strongly. You've all given me strength and determination.

To Maya, who dreams of having her desk next to mine in the postgrad room in "the black hospital" when she grows up- you'll soon discover you don't have to follow my path, you'll get to follow yours. To Myriam, whose entire 7 months of life have been coloured by methylation- you've been so sweetly patient and you don't even know it. To Yazann, for being so understanding and smoothing the wrinkles out from every side. Although you're not a scientist, you know the value of digging deeper until you get to the truth, and we discover what's not a lelandia is a cypress tree. And you're right, this is better than a honeymoon….

Abbreviations

AID	Activation-induced (cytosine) deaminase
AML	Acute myeloid leukaemia
AP	Apurinic/apyrimidinic site
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
APRT	Adenine phosphoribosyl transferase
ATM	Ataxia telangiectasia mutated
BCR	B-cell receptor
BER	Base Excision Repair
BSA	Bovine serum albumin
CBS	Conventional Bisulfite Sequencing
cDNA	Complementary DNA
CGI	CpG island
CHARM	Comprehensive high throughput arrays for relative methylation
ChIP	Chromatin immunoprecipitation
CLL	Chronic lymphocytic leukaemia
CpG	CG dinucleotide
CTCF	CCCTC-binding factor
CTOB	Complementary to original bottom
CTOT	Complementary to original top
Cy3	Cyanine-3
Cy5	Cyanine-5
DHS	DNase hypersensitive site
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTP	2'-deoxynucleotides 5'-triphosphate
DTT	Diothiothreitol
EBV	Ebstein-Barr virus
EC	Embryonal carcinoma
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoretic mobility shift assay
ENCODE	Encyclopedia of DNA Elements
ER	Estrogen receptor
ESC	Embryonic stem cell
FACS	Fluorescence assorted cell sorting
FCR	Fc receptor
HDAC	Histone Deacetylase
HELP	HpaII Tiny Fragment Enrichment by Ligation-mediated PCR
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HERV	Human endogenous retroviruses
HLH	Helix Loop Helix
HPLC	High Performance Liquid Chromatography
ICF	Immunodeficiency, Centromere Instability, Facial anomalies syndrome
IFN	Interferon
iPS	Induced pluripotent stem cells

IGHV	Immunoglobulin Heavy Chain Variable Region
KB	Kilobases
kDa	Kilodaltons
LINE1	Long Intersperesed Nuclear Elements 1
LRES	Long Range Epigenetic Silencing
LTR	Long Terminal Repeats
M	Million
MACS	Magnetic-activated cell separation
MALDI	Matrix Assisted Laser Desorption/Ionisation
MB	Megabases
MBD	Methyl Binding Domain
MBL	Monoclonal B-cell Lymphocytosis
MDS	Myelodysplastic Syndrome
MeDIP	Methylated DNA Immunoprecipitation
Mg	Milligram
Min	Minutes
MIRA	Methylated Island Recovery Assay
miRNA	Micro RNA
Ml	Millilitres
mM	Millimolar
MPN	Myeloproliferative Neoplasms
MSP	Methylation Specific PCR
NER	Nucleotide Excision Repair
Ng	Nanogram
nM	Nanomolar
NLS	Nuclear localization signal
NSCLC	Non-small cell lung cancer
OB	Original bottom
OT	Original top
PARP	Poly ADP ribose polymerase
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate buffered saline
PCNA	Proliferating cell nuclear antigen
PCR	Polymerase chain reaction
PGC	Primordial germ cell
PHD	Plant homeodomain
PLL	Pro-lymphocytic leukaemia
PMSF	Phenylmethanesulfonylfluoride
PMT	Photo-multiplier tube
PTH	Parathyroid hormone
rDNA	Ribosomal DNA
RE	Restriction enzyme
RLGS	Restriction landmark genomic scanning
RNA	Ribonucleic acid
RPA	Random Primer Amplification
RPM	Revolutions per minute
RQ-PCR	Real-time quantitative PCR
RRBS	Reduced representation bisulfite sequencing
RT	Reverse transcriptase

SAM	S-adenosyl methionine
SD	Standard deviation
SDS	Sodium docedyl-sulphate
Seq	Sequencing
SINE	Short interspersed nuclear element
SMRT	Single molecule real time (sequencing)
SNP	Single nucleotide polymorphism
SSC	Saline sodium citrate
TAE	Tris-EDTA buffer
TBE	Tris-borate EDTA buffer
TBS	Targeted Bisulfite Sequencing
TDG	Thymine DNA glycosylase
TOF	Time of Flight
TRD	Transcriptional Repression Domain
TSS	Transcription Start Site
TWEEN	Polyoxyethylene-sorbitan monolaurate
U	Units
UCSC	University of California at Santa Cruz
UTR	Untranslated Region
VNTR	Variable Number Tandem Repeat
WGA	Whole genome amplification
WIMM	Weatherall Institute of Molecular Medicine
WTCHG	Wellcome Trust Centre for Human Genetics
µg	Microgram
µl	Microlitre
µM	Micromolar
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine

Table of contents

CHAPTER 1 . INTRODUCTION **12**
1.1 EPIGENETICS 12
1.1.1 Histones: a structural scaffold and a "code" *13*
1.1.2 DNA methylation: "the 5th base" *14*
1.1.2.1 The modification 14
1.1.2.2 Effects of DNA methylation on genome evolution 15
1.1.2.3 DNA methylation in evolution 16
1.1.2.4 Mammalian patterns of methylation 16
1.2 ESTABLISHMENT AND MAINTENANCE OF DNA METHYLATION 19
1.2.1 DNA methylation through embryonic development *19*
1.2.2 An old base makes a new appearance: 5hmC *20*
1.2.3 The methylation "writers": DNMTs (DNA methyltransferases) *22*
1.2.4 The methylation "readers": MBDs (Methyl Binding Domain proteins) *25*
1.2.5 The methylation "erasers": new players on the scene *27*
1.2.5.1 Demethylation 27
1.2.5.2 Passive demethylation 27
1.2.5.3 Active demethylation 29
1.3 FUNCTIONS OF DNA METHYLATION IN NORMAL CELLS 32
1.3.1 Gene imprinting *32*
1.3.2 Silencing of repetitive elements *32*
1.3.3 Maintenance of genome integrity *34*
1.3.4 Regulation of gene expression *34*
1.3.5 Methylation and structure *37*
1.4 CROSS-TALK BETWEEN DNA METHYLATION AND HISTONE MODIFICATIONS 38
1.4.1 When Histone modifications orchestrate *38*
1.4.2 When DNA methylation directs *39*
1.5 EPIGENETIC ALTERATIONS IN MALIGNANCY 41
1.5.1 Global DNA hypomethylation *41*
1.5.2 Methylation at repetitive elements *42*
1.5.3 Localised hypermethylation *43*
1.5.4 Away from the islands and towards the shores, and back again *45*
1.5.5 The epigenetic machinery in cancer *46*
1.5.5.1 DNA methyltransferases 46
1.5.5.2 Histone proteins 48
1.5.5.3 MBD proteins 49
1.5.6 Are epigenetic mechanisms recapitulating stem cell epigenetics? *49*
1.5.7 Towards a new outlook on methylation anomalies in malignancies *51*
1.5.8 Clinical implications of DNA methylation *53*
1.6 TECHNIQUES FOR STUDYING DNA METHYLATION 54
1.6.1 Methylation typing *55*
1.6.1.1 Methylation-sensitive restriction enzyme analysis 55
1.6.1.2 Bisulfite sequencing 56
1.6.1.3 Pyrosequencing 57
1.6.1.4 MALDI-TOF (Matrix-assisted laser desorption ionization time-of-flight) 57
1.6.1.5 Methylation-specific PCR 58
1.6.2 Methylation profiling *58*
1.6.2.1 RLGS (restriction landmark genomic scanning) 58
1.6.2.2 RRBS (reduced representation bisulfite sequencing) 59
1.6.2.3 MeDIP (Methylated DNA Immunoprecipitation) 59
1.6.2.4 CHARM (comprehensive high throughput array for relative methylation analysis) 60
1.6.2.5 MIRA (methylated island recovery assay) 60
1.6.2.6 Illumina HumanMethylation 27 BeadChip 61
1.6.3 A new gold standard *61*
1.6.4 The next horizon *62*
1.7 CHRONIC LYMPHOCYTIC LEUKAEMIA 63
1.7.1 The clinical condition *63*

1.7.2 Genetic anomalies in CLL *64*
1.7.3 Epigenetic anomalies in CLL *64*
1.8 OUTSTANDING QUESTIONS IN THE FIELD 67
1.9 PROJECT PLAN 69

CHAPTER 2. MATERIALS AND METHODS 71
2.1 ETHICAL APPROVAL 71
2.2 CHEMICALS AND REAGENTS 71
2.3 PRIMERS AND ANTIBODIES 71
2.4 HUMAN TISSUE SAMPLES 71
2.4.1 Normal volunteers *71*
2.4.2 CLL samples *71*
2.4.3 Cell isolation *71*
2.4.3.1 Peripheral blood mononuclear cells (PBMC) 71
2.4.3.2 Neutrophils 72
2.4.3.3 White cell differential isolation by MACS 72
2.4.3.4 CD19 cell selection by DynaBeads 72
2.4.4 Assessment of cell purity *73*
2.4.4.1 FACS 73
2.4.4.2 Cytospins 73
2.5 TISSUE CULTURE CELL LINES 73
2.6 DNA PROCEDURES 73
2.6.1 Extraction and quantitation of DNA *73*
2.6.1.1 From cell pellets 73
2.6.1.2 From bacterial cultures (MiniPreps) 74
2.6.2 Polymerase Chain Reaction (PCR) *74*
2.6.3 Quantitative real-time PCR *74*
2.6.3.1 For ChIP assessment 74
2.6.3.2 For protein expression analysis 75
2.6.3.3 For MeDIP- Diagenode kit 75
2.6.3.4 For MEDIP- VR protocol 75
2.6.3.5 For targeted capture protocol 75
2.6.4 DNA sequencing *76*
2.6.4.1 Sanger sequencing 76
2.6.4.2 High throughout sequencing 76
2.6.5 DNA digestion with restriction enzymes *76*
2.6.6 Southern blotting *77*
2.6.7 DNA ligation *77*
2.7 RNA PREPARATION AND ANALYSIS 78
2.7.1 Extraction and quantitation of RNA *78*
2.7.2 cDNA synthesis and reverse transcription PCR *78*
2.8 PROTEIN PREPARATION AND ANALYSIS 79
2.8.1 Nuclear extracts *79*
2.8.2 Quantitation of protein *79*
2.8.3 Coomassie and silver stain *79*
2.8.4 Electro Mobility Shift Assays (EMSAs) *80*
2.8.5 DNA affinity chromatography *81*
2.8.5.1 Preparation of double-stranded oligos 81
2.8.5.2 Preparation of columns 81
2.8.5.3 Blocking of the beads 81
2.8.5.4 Preparation of the extracts 82
2.8.5.5 Incubation and washes 82
2.8.5.6 Elution 82
2.8.6 MALDI-TOF *82*
2.9 CHROMATIN PREPARATION AND ANALYSIS 83
2.9.1 Chromatin immunoprecipitation *83*
2.9.2 Labelling and hybridisation to microarrays *83*
2.10 METHYLATED DNA ANALYSIS 84

2.10.1 Bisulfite sequencing ... *84*
2.10.1.1 Bisulfite conversion ... 84
2.10.1.2 Nested PCR ... 85
2.10.2 MeDIP (methylated DNA immunoprecipitation) ... *85*
2.10.2.1 Weber protocol ... 85
2.10.2.2 VR protocol ... 86
2.10.2.3 Diagenode protocol ... 86
2.10.2.4 Whole Genome Amplification (WGA) ... 87
2.10.2.5 Labelling/hybridising/scanning ... 87
2.10.3. Targeted capture ... *88*
2.10.3.1 Sonication ... 88
2.10.3.2 Capture ... 88
2.10.3.3 Bisulfite conversion ... 89
2.10.3.4 Random Primer Amplification (RPA) ... 89
2.10.3.5 AcuI digestion ... 90
2.10.3.6 Gel purification ... 90
2.10.4 HPLC analysis for 5hmC levels ... *90*
2.11 BACTERIAL PROCEDURES ... 91
Transformation of competent cells ... *91*
2.11 BIOINFORMATICS ... 91
2.13 STATISTICS ... 92

CHAPTER 3. DNA METHYLATION ABNORMALITIES IN CLL ... 93
3.1 INTRODUCTION ... 93
3.2 PRELIMINARY FINDINGS ... 93
3.2.1 Methylation abnormalities at 16p in malignancies ... *93*
3.2.3 DNA methylation in CLL at repetitive elements ... *95*
3.3 CONVENTIONAL BISULFITE SEQUENCING (CBS) ... 98
3.3.1 Overview of the technique ... *98*
3.3.2 Technical considerations ... *98*
3.3.3 Choice of samples ... *102*
3.3.3.1 CLL samples ... 102
3.3.3.2 Normal samples ... 104
3.3.4 Changes in DNA methylation in CLL ... *109*
3.3.4.1 Technical details ... 109
3.3.4.2 Direction and magnitude of changes from normal to CLL ... 109
3.3.4.3 Distribution of changes ... 109
3.3.4.4 Motif analysis ... 111
3.3.4.5 Extension to further samples ... 113
3.4 METHYLATED DNA IMMUNOPRECIPITATION (MEDIP) ... 113
3.4.1 Method overview and rationale ... *113*
3.4.2 Establishment of a protocol ... *117*
3.4.2.1 DNA preparation ... 117
3.4.2.2 Immunoprecipitation ... 117
3.4.2.3 Whole Genome Amplification (WGA) ... 122
3.4.2.4 Labelling/hybridising/scanning ... 123
3.4.2.5 Data manipulation and interpretation ... 123
3.4.3. Application to an experiment ... *128*
3.4.3.1 Experimental design ... 128
3.4.3.2 Quality control ... 128
3.4.3.3 Comparison between MeDIP and bisulfite data ... 128
3.4.3.4 Differences in methylation between normal and CLL ... 131
3.4.3.5 CpG islands ... 134
3.5 DISCUSSION ... 136

CHAPTER 4. HIGH RESOLUTION ASSESSMENT OF EPIGENETIC CHANGES IN CHROMOSOME 16P ... 139
4.1 INTRODUCTION ... 139
4.2 TARGETED BISULFITE SEQUENCING (TBS) ... 141
4.2.2 Establishment of a protocol ... *141*

4.2.2.1 Overview of the protocol 141
4.2.2.2 Optimisation of individual steps 144
4.2.3 High Throughput sequencing 148
4.2.4 Data manipulation 148
4.2.4.1 Fast QC 150
4.2.4.2 Bismark 150
4.2.4.3 Methylation Extractor 153
4.2.5 Initial results: technical issues 154
4.2.5.1 Fast QC results 154
4.2.5.2 Poor mapping efficiency 156
4.2.5.3 Inaccurate mapping 156
4.2.5.4 PCR bias 158
4.2.5.5 Strand discrepancy 158
4.2.5.6 CpG density bias 161
4.2.6 Validation of the final protocol 162
4.2.6.1 Correlations between Conventional and Targeted Bisulfite Sequencing 164
4.2.6.2 Examination of individual areas 164
4.2.6.3 Identification of differentially methylated regions between normal and CLL 164
4.2.7 Results: Methylation changes in CLL patients 166
4.2.7.1 Sequencing and mapping results 166
4.2.7.2 Direction and magnitude of changes from normal to CLL 167
4.2.7.3 Distribution of changes 169
4.2.7.4 Extreme changes in methylation 172
4.2.7.5 Motif analysis 172
4.3 HISTONE MODIFICATIONS 177
4.3.1 Rationale 177
4.3.2 ChIP-on-chip 177
4.3.3 ChIp-Seq 178
4.4 DISCUSSION 184
4.4.1 Technical approach 184
4.4.2 Summary of epigenetic results at 16p 185
4.4.2.1 DNA methylation 185
4.4.2.3 Histone modifications 188
4.4.3 Elaboration of a hypothesis 188

CHAPTER 5. PASSIVE HYPOMETHYLATION 191
5.1 INTRODUCTION 191
5.2 PRELIMINARY STUDIES 193
5.2.1 Enzyme levels 193
5.2.2 5hmC levels 195
5.2.3 Choice of area of study 197
5.2.4 Candidate proteins 200
5.3 IDENTIFICATION OF PROTEINS IN NUCLEAR EXTRACTS FROM EMSAS 203
5.3.1 Samples/protocol 204
5.3.1.1 CLL samples 204
5.3.1.2 Normal samples 204
5.3.1.3 Nuclear extracts 205
5.3.1.4 Gel shift protocol 205
5.3.2 Screening gel shifts 207
5.3.3 Primary CLLs 211
5.3.4 Methylated oligo 211
5.4 DNA AFFINITY CHROMATOGRAPHY 211
5.4.1 Mutant oligos 215
5.4.2 Optimisation of the DNA affinity chromatography protocol 218
5.4.2.1 Outline of the protocol 218
5.4.2.2 Plan to test the protocol 218
5.4.2.3 Initial pull-down 220
5.4.2.4 Varying buffers 220
5.4.3 Further optimisation 222
5.4.3.1 Second pull-down with altered buffers 222

5.4.3.2 Determination of possible causes 222
5.4.4 Silver stain *224*
5.4.5 MALDI-TOF results *226*
5.5 DISCUSSION AND FURTHER WORK 228
5.5.1 Pursuing the silver stain and MALDI-TOF avenue *228*
5.5.2 Trying to pursue the use of EMSAs as a monitoring assay *229*
5.5.3 Optimising the original protocol *230*
CHAPTER 6. FINAL DISCUSSION AND CONCLUSIONS 231
APPENDIX 1: PRIMERS AND ANTIBODIES 237
APPENDIX 2: SUPPLEMENTARY FIGURES 244
REFERENCES 251

Chapter 1. Introduction

"Now Equilibrium is the very opposite of Disorder." –Rudolf Arnheim

Every organism must exist in a state of equilibrium between opposing forces. For example, a delicate balance must exist between tight control of gene expression on one side and sufficient plasticity to allow for adaptation to a changing environment on the other. Cells which manage to sufficiently tip the balance away from a situation of tight control can escape into unbridled proliferation and towards the ultimate disordered state for the organism, malignancy. The layers of control over gene expression are exquisitely varied and cooperate in intricate manners to ensure this equilibrium is maintained over the course of the organism's life. DNA methylation, a modification to one of the canonical bases of the genetic code, plays a major, if at times still elusive, role in this process, and its derangement is a hallmark of malignancy. Clarifying the mechanisms for its anomalies in malignancy may answer critical questions, both regarding how genes are normally regulated, and how malignant cells can escape the normal orchestration to which the genome is subjected.

1.1 Epigenetics

The term "epigenetics" has evolved since its first introduction in the literature by Waddington in 1954 and is generally accepted to mean "heritable changes in gene function not resulting from changes in DNA sequence". It refers to the processes by which distinct cell types can express unique transcription programmes despite containing an identical genome. Epigenetics encompasses the layers of regulation that control the genetic information and thus govern key processes such as embryonic development, lineage differentiation, and control of gene expression to name a few. As our understanding of the mechanisms that mediate epigenetic processes involved expands, so does the list of the different aspects that are comprised under this broad umbrella that is "epigenetics".

Nevertheless, two key components of epigenetic regulation are histone modifications and DNA methylation.

1.1.1 Histones: a structural scaffold and a "code"

One turn of the DNA double helix winds around octamers of histone proteins to form nucleosomes. The term "chromatin" refers to the complex of DNA and its associated histone proteins. The canonical histones that form the basic nucleosome are H2A, H2B, H3, and H4, but much interest has focused recently on histone variants. For example histone H3.3 is found at sites of transcriptionally active genes (HENIKOFF and AHMAD 2005; KOUZARIDES 2007). As such, this is an example of an epigenetic modification where gene activity is associated with a change in proteins and not a change in DNA sequence. Furthermore, the structure of the histone proteins makes them eminently suited to their role of harbouring signals for control of DNA. They contain a structured globular domain and a long unstructured N-terminal tail. These tails are subjected to an ever-expanding list of ways in which certain residues may be covalently modified. Common modifications involve methylation, acetylation, ubiquitylation, phosphorylation, or sumoylation and to date over 60 have been described (KOUZARIDES 2007). Modifications are most often found at lysine (K) residues, although arginine (R) and serine (S) are also targets for modifications. Some residues can be modified multiple times such as mono-, di- or tri-methylation (KOUZARIDES 2007). The presence, location and combination of these different modifications exert complex regulatory effects on the surrounding DNA. This is sometimes a direct effect, such as lysine acetylation, which carries a negative charge, repelling the tails from surrounding DNA and rendering the DNA more accessible to soluble factors. Alternatively, histone modifications may have indirect effects through "readers" of these marks, such as the H3K27me3 mark laid down by the polycomb complex PRC2, attracting binding of another repressive polycomb complex, PRC1, which seems to cause chromatin compaction by as yet unclear mechanisms (FRANCIS *et al.* 2004).

Traditionally, histone marks have been divided into "active" (such as H3Ac, H4Ac, H3K4me3), and "repressive" (H3K27me3, H3K9me3) marks, although this

distinction may not be absolute in reality. For example, H3K9me3, a typical repressive mark, has now been shown to increase with transcriptional elongation (HAHN *et al.* 2011; VAKOC *et al.* 2005). Some of the modifications are important in marking different regions in the genome, such as promoters or enhancers, with the former characterised by H3K4me3 and the latter by H3K4me1 and H3K27ac in their active state (BAYLIN and JONES 2011).

Histones are generally inhibitory to transcription as they shelter DNA from transcription factor access, and the RNA polymerase responsible for mRNA synthesis from DNA cannot transcribe through nucleosomes (GILBERT and RAMSAHOYE 2005). Therefore, some of the transcriptional control exerted by histones is not due to the histone variants or their modifications but by their absence from a locus, such as a lack of nucleosomes marking transcription start sites (reviewed in (KOUZARIDES 2007). A complete lack of nucleosomes allows unfettered access of transcription factors to DNA, and is reflected experimentally by the presence of DNAse hypersensitive sites. These have unequivocally been shown to appear at loci of active chromatin in a timing consistent with their role in the activation process (ELGIN 1988).

The way in which different histones and their modifications influence the transcription of DNA has been dubbed "the histone code" (JENUWEIN and ALLIS 2001). In contrast with this complex system, the other main component of epigenetics, DNA methylation, had seemed comparatively simple due to the involvement of a single base modification, and yet now appears to be no less intricate in its regulating potential.

1.1.2 DNA methylation: "the 5th base"

1.1.2.1 The modification

For numerous years, methylation was described as the only known covalent modification of DNA, and in eukaryotes it is found exclusively at cytosine residues. It is created by the transfer of a methyl group from S-adenosyl methionine (SAM) to the carbon-5 position of cytosine by DNA methyltransferases (which will be discussed in section 1.2.3). In plants,

methylation occurs most commonly in a CpG context (C followed by G), although it is less frequently seen in a CHG context (H=A, T or G), and even more rarely in a CHH context (COKUS *et al.* 2008). In mammals, methylation was thought to occur almost exclusively in the context of a CpG dinucleotide, with very rare exception (CLARK *et al.* 1995). It was later shown that non-CpG methylation does occur in mammals, particularly in ES cells, where 15-20% of cytosine methylation is in a non-CpG context (RAMSAHOYE *et al.* 2000). This finding has only recently been confirmed and extended by high throughput whole genome bisulfite sequencing which shows that in ES cells, only 75.5% of methylation is in a CpG context, with 17.3% at CHG and 7.2% at CHH (LISTER *et al.* 2009) sites. CpG methylation normally occurs symmetrically such that the complementary C to the G is also methylated. Where this does not occur, it is termed "hemi-methylation", a concept which will be alluded to later.

1.1.2.2 Effects of DNA methylation on genome evolution

Methylated CpGs are prone to spontaneous deamination, the product of which is a thymine (T). This creates a T:G mismatch which is not always recognised by the DNA repair machinery. The result is twofold: depletion of CpG dinucleotides from the genome, and secondly, mutagenesis (COOPER and KRAWCZAK 1989). As a result, CpGs are vastly under-represented in the genome, being found at only 25% of the expected rate (SAXONOV *et al.* 2006). While a random distribution would predict that CpGs would be present at 1:16 nucleotides, they are only found at 1:80 dinucleotides (LEHMANN *et al.* 2004). Furthermore, these are unevenly distributed. Ninety-eight percent of CpGs are scattered throughout the genome at low densities, while the remaining 2% are found clustered in high-density regions termed CpG islands (CGIs) (ILLINGWORTH *et al.* 2008). These mark the promoters of ~70% of protein-coding genes, encompassing nearly all housekeeping genes and ~40% of tissue-specific genes (ILLINGWORTH and BIRD 2009). These two classes of CpGs, as based on their distribution, are subject to very different methylation patterns. Whereas the isolated CpGs, (found mostly in repetitive sequences or gene bodies) are densely methylated, CGIs tend to be unmethylated (SUZUKI and BIRD 2008).

The second consequence of spontaneous CpG deamination, mutagenesis, has interestingly been proposed to be responsible for the creation of multiple new transcription factor binding sites in the genome, particularly at highly repetitive sequences termed *Alu* repeats, thereby increasing gene expression variability on which the selective pressures of evolution can be applied (ZEMOJTEL *et al.* 2011).

1.1.2.3 DNA methylation in evolution

While the DNA methyltransferases are highly conserved throughout evolution (BESTOR 2000), the patterns of methylation do vary between different species. Drosophila have trace amounts of methylation (0.05-0.1% of cytosine residues) (GOWHER *et al.* 2000), while some species, e.g. *S. cerevisiae* and *C. elegans* are devoid of methylation. Bacteria use the patterns of DNA methylation to recognise foreign DNA and target it for degradation by restriction endonucleases. As reviewed in Suzuki *et al.* (SUZUKI and BIRD 2008) *N. crassa* and plants have mosaic DNA methylation with a mostly unmethylated genome interspersed with densely methylated stretches corresponding to transposable elements. In addition, plant genomes are characterised by methylation in the bodies of genes. The pattern is very different in vertebrates, where most CpGs are densely methylated (corresponding to gene bodies and transposable elements), and unmethylated CpGs found clustered in islands. Within gene bodies, eukaryotes have higher methylation in exons than introns (LEE *et al.*). Several studies on the patterns of DNA methylation in evolution have concluded that gene body methylation is highly conserved and predates the divergence of plants and animals. Given that methylation of repetitive elements is more variable, this may have evolved in plants and vertebrates independently but not evolved in invertebrates (FENG *et al.* ; ZEMACH *et al.*).

1.1.2.4 Mammalian patterns of methylation

Our understanding of the normal patterns of CpG methylation has grown with the increasing sophistication and precision of the techniques available to study this phenomenon. Early observations based on the differential sensitivity of methylated and unmethylated CpGs to restriction enzyme cleavage led to the identification of the CpG islands mentioned above, which were shown to be unmethylated (BIRD *et al.* 1985). High Performance Liquid Chromatography

(HPLC) analysis of hydrolyzed DNA showed that the level of methylation varies between human tissue types (EHRLICH *et al.* 1982). Across different tissue types, however, DNA that was enriched for repetitive sequences such as *Alu* elements was found to be consistently densely methylated (GAMA-SOSA *et al.* 1983b).

Recent whole-genome bisulfite sequencing studies have helped refine the concept of "gene body methylation" described above. It has now been shown that CpG methylation dips sharply at transcription start sites and rises again in gene bodies (LAURENT *et al.* 2010; LISTER *et al.* 2009). Furthermore there is a sharp spike of DNA methylation at the beginning of a splice junction followed by a deep dip at the end of the splice junction and methylation is higher in exons than introns. It has therefore been proposed that DNA methylation in concert with histone modifications may play a role in the recognition of intron-exon boundaries and be involved in the splicing process (LAURENT *et al.* 2010).

Methylation patterns are segregated into, roughly, unmethylated CGIs and densely methylated genes and intergenic regions. It is therefore not surprising that results from several different techniques-including MeDIP (Methylated DNA Immunoprecipitation) (WEBER *et al.* 2005), targeted bisulfite sequencing (HODGES *et al.* 2009) and reduced representation bisulfite sequencing (GU *et al.* 2010)-have shown that the methylation status of a CpG is highly correlated with the methylation of CpGs in a 1kb surrounding region. This has been found for unmethylated and highly methylated regions, but also for partially methylated regions (HODGES *et al.* 2009). It has also been suggested that the methylation patterns are not replicated through cell division at an individual CpG level but that the methylation machinery ensures that the "methylation entropy" is faithfully replicated. This term has been defined as "the variability of DNA methylation pattern that might be observed for a given genomic locus in a cell population" (XIE *et al.* 2011). Indeed, hairpin bisulfite sequencing studies looking at hemi-methylation have shown that the fidelity of copying exact methylation patterns is about 96% for a hypermethylated island but only 88% for LINE1 elements, despite the overall methylation levels staying constant (LAIRD *et al.* 2004b).

The tissue specificity of methylation patterns has now been described extensively. Some cell types have extremely different methylation patterns from others (e.g. 20% difference between sperm and fibroblasts), whereas two very similar cell types such as CD4 and CD8 cells have <5% differences in methylation (ECKHARDT *et al.* 2006a). This has led to the concept of t-DMRs (tissue-specific differentially methylated regions), first defined by Imamura *et al.* (IMAMURA *et al.* 2001). The patterns of methylated vs unmethylated regions in individual cells can now be used to accurately classify their tissue of origin (ECKHARDT *et al.* 2006a). It is noteworthy that the CGIs which harbour differential methylation between tissues are highly enriched for genes involved in development (IRIZARRY *et al.* 2009). While it remains true that CGIs are on the whole unmethylated, this phenomenon is restricted to CGIs at promoters and transcription start sites (TSSs), whereas intragenic CGIs are densely methylated (HODGES *et al.* 2009). Strikingly, CGIs involved in differential methylation between tissues are more commonly those away from TSSs (22%) than within them (8%) (ILLINGWORTH *et al.* 2008).

While methylation patterns are relatively uniform within a cell type (ECKHARDT *et al.* 2006a), some variation between individuals occurs. In humans, overall methylation has been reported to decrease by 9% incrementally with each decade of life (YU *et al.* 2007). There is more variation in DNA methylation between older monozygotic twin pairs than younger ones, (FRAGA *et al.* 2005a), suggesting an environmental component to determining DNA methylation patterns. Studies in rats have demonstrated an accumulation of changes in DNA methylation patterns in tissues, and the changes are more marked in rapidly dividing cells than quiescent ones (THOMPSON *et al.* 2010). Other studies focusing on the methylation of specific regions of three chromosomes did not, however, find age or sex-related differences in methylation (ECKHARDT *et al.* 2006a). Interestingly, analysis of polymorphisms found in the DNMT genes identified a SNP in DNMT3L which is associated with hypomethylation in subtelomeric regions (EL-MAARRI *et al.* 2009), suggesting a genetic mechanism for methylation

differences between individuals in addition to presumed environmental influences that would be inferred from the studies on monozygotic twins.

1.2 Establishment and maintenance of DNA methylation

1.2.1 DNA methylation through embryonic development

The pattern of DNA methylation is established in a complex manner, the control of which remains poorly understood. While mature gametes have an established methylation pattern, this is erased in the pre-implantation embryo in a rapid wave of demethylation. It is quickly re-established at implantation, when most CpGs are re-methylated apart from those in CGIs (reviewed in (WEISS and CEDAR 1997). The protection of CGIs from *de novo* methylation has long been a mechanistic puzzle. It was initially reported that the transcription factor Sp1 may play a role in protecting CGIs from methylation as it was shown in a transgenic mouse model system that Sp1 binding sites were necessary to prevent a CGI in the *APRT* gene from becoming methylated (BRANDEIS *et al.* 1994). Furthermore, Sp1 has been shown to protect DNA from hypermethylation in malignancy in the K562 (human erythroleukaemia) cell line (BOUMBER *et al.* 2008). Recent evidence linking the protection of CGIs from methylation to the presence of methylated histones at these locations has also been put forward. RNA polymerase II, which is bound to CGIs in the early embryo, recruits a specific histone methyltransferase which methylates the lysine 4 residue on histone H3. This mark excludes DNMT3L, a cofactor responsible for recruiting DNMT3A and 3B (OOI *et al.* 2007). As a result, it is proposed that CGIs are not subjected to the global wave of remethylation that occurs at this stage.

The methylation pattern in human ES cells (hES) has become clearer recently. The high proportion of non-CpG methylation (25%) in these cells, although now well described, plays a role which is still under investigation. Intriguingly, 98% of non-CpG methylation is found on one strand only, with a highly statistically significant preference for the antisense strand to the gene body, the significance of which has not been addressed (LISTER *et al.* 2009). Furthermore, particular genes in hES cells are hypomethylated compared to differentiated cells. These

include developmentally regulated genes, in particular the *HOX* gene clusters (LAURENT *et al.* 2010). Differentially methylated regions between hESCs and hESC-derived fibroblasts are associated with genes involved in pluripotency and development, and, as cells differentiate, they gain methylation at homeobox transcription factors, raising the possibility that DNA methylation plays a role in specifying transcriptional programmes involved in lineage commitment (LAURENT *et al.* 2010).

The study of methylation in embryonic stem cells has been extended to include methylation in induced pluripotent stem cells (iPS cells). When fibroblasts are reprogrammed into iPS cells, they regain a methylation pattern analogous to that in embryonic cells. This, however, is imperfect, as some sites fail to get reprogrammed while others are aberrantly reprogrammed (DOI *et al.* 2009). Importantly, genes which lose methylation on reprogramming include key pluripotency genes such as *OCT4*, *Nanog*, as well as *DNMT3B*, all of which are highly expressed in stem cells. In parallel, 80% of promoters with "bivalent" chromatin marks in ES cells regain these in iPS cells (DE CARVALHO *et al.* 2010). Overall, iPS cells are hypermethylated compared to embryonic stem cells, which may be due to the continued over-expression of viral transgene-driven transcription factors (DENG *et al.* 2009).

1.2.2 An old base makes a new appearance: 5hmC

Although it was originally thought that methylation was the only covalent modification of DNA, this view has been challenged by recent work on other modifications such as 5-hydroxymethylcytosine (5hmC). This modified base was first identified in the 1950s (WYATT and COHEN 1952) where it was thought to be a uniquely viral base. Its meteoric rise in importance occurred following the discovery that it is the previously unidentified base present in Purkinje neurons (KRIAUCIONIS and HEINTZ 2009). It was simultaneously reported that 5mC can be converted to 5hmC by TET (Ten-eleven translocation) enzymes. TET1 was identified from a search for enzymes able to modify 5mC. This search originated from trypanosome 2(oxoglutarate) and Fe(II) dependent oxygenases, which modify thymine into J-bases, a trypanosome-specific modification involved in the

expression of host defence evasion genes. Sequence homology was sought in mammalian genomes and TET1, TET2 and TET3 identified (TAHILIANI *et al.* 2009). These enzymes were shown to cause reduction of 5mC when over-expressed in HEK293 (human embryonic kidney) cells and caused the appearance of a new species of nucleotide on thin layer chromatography. Mass spectrometry of this nucleotide identified it as 5hmC, and TET1 was shown to convert 5mC to 5hmC through hydroxylation, requiring Fe(II) and 2OG.

The genome-wide distribution of 5hmC has recently been determined in ES cells where it is found enriched at exons and transcriptional start sites (PASTOR *et al.* 2011). It is particularly enriched at promoters with bivalent chromatin marks, however its role is still a matter of conjecture. It appears to play an important role in ES cell self-renewal where 5hmC is crucial for expression of *Nanog*. In these experiments, TET1 knock-down led to *Nanog* hypermethylation and loss of expression (ITO *et al.* 2010). TET1 and TET2 are both vitally important in ES cell self-renewal and cell lineage specification and can both convert 5mC to 5hmC (ITO *et al.* 2010), (KOH *et al.* 2011).

The importance of the TET enzyme family is highlighted by its involvement in haematological malignancies. *TET1* is a known fusion partner with *MLL* in AML (LORSBACH *et al.* 2003) and *TET2* mutations have been identified in MDS, MPN, and AML (DELHOMMEAU *et al.* 2009; TEFFERI *et al.* 2009). Furthermore, a loss of *TET2* in a conditional mouse knockout leads to increased haematopoietic stem cell self-renewal and a myeloproliferative bone marrow (MORAN-CRUSIO *et al.* 2011), implicating TET2 in the control of stem cell characteristics.

Finally, other modified bases have since also been reported, such as 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), which can be products of the TET enzymes (ITO *et al.* 2011). 5-carboxylcytosine can be excised by TDG (thymine DNA glycosylase), thereby potentially playing a role in active DNA demethylation as will be explained below (HE *et al.* 2011; MAITI and DROHAT 2011).

1.2.3 The methylation "writers": DNMTs (DNA methyltransferases)

The enzymes with DNA methylation activity are termed DNMTs and include DNMT1, DNMT3A, and DNMT3B. In the 24 years since DNMT1 was cloned (BESTOR *et al.* 1988), followed by DNMT3A and 3B a decade later (OKANO *et al.* 1998), our understanding of the roles of these enzymes has changed. Since DNMT1 has a much greater affinity for hemi-methylated CpGs than for unmethylated ones, its role as a maintenance methyltransferase was deemed logical. This hypothesis was strengthened by several key findings: DNMT1 is targeted to replication foci (LEONHARDT *et al.* 1992); it is most highly expressed in S-phase when DNA replication is occurring and methylation of the daughter strand is necessary (SZYF *et al.* 1991); and finally it associates with PCNA ((CHUANG *et al.* 1997) and UHRF1 (BOSTICK *et al.* 2007) which are found at the replication fork and therefore ideally suited to target DNMT1 to sites requiring methylation. Disruption of UHRF1 leads to loss of DNA methylation at repetitive elements and individual gene loci (BOSTICK *et al.* 2007). Furthermore, UHRF1 plays a role in recognizing hemi-methylated DNA, leading to a base-flipping mechanism exposing the unmethylated CpG to DNMT1 (ARITA *et al.* 2008). DNMT1 has recently been shown to be excluded from unmethylated CpGs by an auto-inhibitory effect through its CXXC domain. When DNMT1 binds unmethylated CpGs, the CXXC domain inhibits the active site, ensuring only hemi-methylated CpGs are targeted by the enzyme at the replication fork (SONG *et al.* 2011).

The DNMT3 family comprises DNMT3A, DNMT3B and DNMT3L. The latter does not contain a catalytic domain itself but is necessary for the function of both DNMT3A and 3B (SUETAKE *et al.* 2004) and also seems to play a role in imprinting at maternal loci (BOURC'HIS *et al.* 2001). It does not alter the sequence preference of DNMT3A or 3B but in an *in vivo* episomal context, targets them to well-chromatinised templates and increases their efficiency at sites where this is particularly low (WIENHOLZ *et al.* 2010).

DNMT3A and 3B can equally well methylate unmethylated and hemi-methylated CpGs (OKANO *et al.* 1998). They are most highly expressed in ES cells and

downregulated (but still expressed) in somatic cells (CHEN *et al.* 2003). Mice lacking DNMT3A and 3B do not survive beyond 4 weeks (OKANO *et al.* 1999). These enzymes were therefore deemed to be the *de novo* methyltransferases, responsible for setting up the DNA methylation patterns in the embryo, with DNMT1 then being responsible for maintenance methylation thereafter. Consistent with this, ES cells do not need DNMT1 to grow but differentiated cells in which DNMT1 has been knocked out die by a p53-dependent mechanism (JACKSON-GRUSBY *et al.* 2001).

Despite the seemingly neat separation of roles for DNMT1 and DNMT3A and 3B, evidence has accumulated that this explanation was over-simplistic. Furthermore, in a provocative review, Jones ((JONES and LIANG 2009) highlighted that there seemed to be no mechanism for proofreading of the methyl marks in this system.

The first line of evidence against DNMT1 being the sole enzyme responsible for maintenance methylation was that it was evident that there was poor fidelity in maintaining methylation (POLLACK *et al.* 1980; WIGLER *et al.* 1981), and it has been estimated that there is a loss of 5% methylation per CpG site per cell division (BIRD 2002). Furthermore, mouse ES cells that contain DNMT1 but not DNMT3A or 3B gradually lose methylation with successive cell divisions (CHEN *et al.* 2003; HATTORI *et al.* 2004). This implies that DNMT3s are somehow involved in the maintenance of methylation. In particular, when DNMT3A and 3B are knocked down, 30% of CpGs in repeats in the mouse genome become hemi-methylated, suggesting the importance for these enzymes in maintenance of methylation particularly at these sequences ((LIANG *et al.* 2002).

Conversely, DNMT1 seems to play a role in *de novo* methylation. In an *in vitro* system, its addition to DNMT3A increases methylation efficiency 5-fold (FATEMI *et al.* 2002).

In the revised model proposed by Jones, the bulk of DNA methylation in a dividing cell is maintained by DNMT1 in conjunction with UHRF1. DNMT1 may perform error correction after it has left the replication fork. For highly methylated sites such as repeats and methylated CGIs, however, DNMT1 may not be able to maintain an adequate rate of methylation and DNMT3A and 3B may be necessary for *de novo* re-methylation at these sites. As DNMT3A and 3B have been shown to localise to areas of high DNA methylation (SCHLESINGER *et al.* 2007) this may ensure that adequate levels of 5mC are maintained at these areas, akin to the finding that the histone methyltransferase G9A remains at repetitive elements to ensure constant H3K9 trimethylation (IKEGAMI *et al.* 2007).

While DNMT1 is targeted to sites of hemi-methylation by UHRF1, how DNMT3A and 3B are targeted to unmethylated sites requiring methylation is still unclear. They have been shown to localise to pericentric heterochromatin in embryonic stem cells through their PHD domains (BACHMAN *et al.* 2001). Bioinformatic analyses have identified strong sequence preferences for recruitment and exclusion of DNMT3A and 3B, and these sequences have been validated by *in vitro* experiments. This would seem to explain some of the methylation hotspots found in the genome ((HANDA and JELTSCH 2005). DNMT1 may also be targeted more specifically. For example, it has been shown to be part of a complex that targets HDACs to *E2F* promoters, suggesting DNMT1 may cause methylation to participate in repression at those loci (ROBERTSON *et al.* 2000). A search for transcription factors that interact with recombinant DNMT1 identified 58 that are capable of binding the enzyme *in vitro* (HERVOUET *et al.* 2010b). This opens the possibility that the specific targeting of DNMT1 to individual gene loci is a widespread phenomenon.

While DNMT3A and 3B have some redundant functions, they also have some distinct CpG targets (OKA *et al.* 2006). For example, DNMT3A is particularly important at imprinted loci and interspersed repeats whereas DNMT3B is important at pericentromeric satellite repeats (CHEN *et al.* 2003; KANEDA *et al.* 2004; KATO *et al.* 2007).

DNMT3A and 3B are not found free in the nucleus but are strongly bound to some nucleosomes, in particular nucleosomes associated with fully methylated LINEs, SINEs and CGIs (JEONG *et al.* 2009b). Unbound DNMT3A and 3B are rapidly degraded. Furthermore, a reduction in global DNA methylation from DNMT1/DNMT3B double knockouts leads to a reduction in DNMT3A protein levels despite increased mRNA levels. This is due to reduced protein stability from lack of stabilization afforded from binding methylated DNA (SHARMA *et al.* 2011).

1.2.4 The methylation "readers": MBDs (Methyl Binding Domain proteins)

A key question regarding the function of DNA methylation is how the information embedded in the CpG methylation status can be relayed to downstream effector processes. The discovery of Methyl Binding Domain (MBD) proteins, and their direct link with proteins known to modify histone tails has provided the beginning of an understanding of the mechanisms involved. Five MBD proteins have been discovered in mammals and they share a common Methyl Binding Domain (~70bp) that is absolutely required for binding to 5mC. The first MBD protein to be characterised was MeCP2 (MEEHAN *et al.* 1992), and a subsequent search for related proteins identified MBD1-4 (HENDRICH *et al.* 1999a). In addition to its MBD domain, MBD4 contains DNA n-glycosylase activity and it has a role in DNA mismatch repair. Specifically, MBD4 interacts with a mismatch-repair protein called MLH1 to correct T:G mismatches that arise specifically from spontaneous deamination (HENDRICH *et al.* 1999b).

The MBD domain itself does not provide sequence specificity for binding, although MeCP2 has been shown to have a preference for A/T bases adjacent to the methylated CpGs (KLOSE *et al.* 2005). It has also been suggested that the dissociation constant of the MBD proteins dictate to which CpG islands these bind preferentially, rather than any sequence specificity (FRAGA *et al.* 2003). MBD1 can bind both methylated and unmethylated CpGs (FUJITA *et al.* 2000) while MBD3 does not bind methylated DNA despite its MBD (ZHU *et al.* 2009). MBD3 is a component of the Mi2/NuRD complex which has HDAC and ATPase activities, but its role within the complex is not known (SAITO and ISHIKAWA

2002). The four true MBD proteins all associate with protein partners that are known to mediate transcriptional repression and/or the formation of heterochromatin. It is possible that their association with a varying complement of these proteins provides plasticity in the way each MBD protein can function under different conditions (type of cell, surrounding chromatin environment, etc). There is, however, no evidence that the MBDs function to target different sets of genes and from mouse knockout studies it seems more likely that there is some redundancy in their function (FATEMI and WADE 2006).

The function of MeCP2 serves to illustrate how DNA methylation status is relayed to downstream effectors. MeCP2 is a multidomain protein, with a TRD (transcriptional repressor domain) and a NLS (nuclear localization signal) (HITE *et al.* 2009). The TRD associates with Sin3 and HDAC (JONES *et al.* 1998), both known co-repressor molecules which cause a hypoacetylated chromatin environment, restricting access of transcription factors to their target sites. The role of MeCP2, however, is more complex than simply to effect transcriptional repression at sites of methylated DNA. It has been shown *in vitro* to compact chromatin at areas without DNA methylation and in the absence of Sin3a, and as such is thought to have a role in the regulation of global chromatin structure (GEORGEL *et al.* 2003). It may also be involved in RNA splicing (YOUNG *et al.* 2005). Surprisingly, MeCP2 has been shown to associate with actively transcribed genes far more often than with repressed genes (YASUI *et al.* 2007) raising questions about its functions.

The recognition of methylated DNA is not confined to proteins with an MBD domain, as structurally unrelated proteins such as Kaiso proteins (SASAI *et al.* 2010) and newly identified zinc finger proteins (FILION *et al.* 2006) bind methylated DNA and can repress transcription.

Finally, as will be discussed in section 1.6.2, the discovery of MBD proteins and their ability to bind methylated DNA has been exploited in the development of techniques used for methylation profiling.

1.2.5 The methylation "erasers": new players on the scene

1.2.5.1 Demethylation

While the enzymes involved in DNA methylation have been known for twenty years, the process by which demethylation occurred, if it did, remained a mystery until recently. It was initially thought that DNA methylation was a very stable epigenetic modification in adult somatic tissues (LIN and HSIEH 2001) and there was debate as to whether demethylation even occurred. Early hypotheses considered demethylation to occur as a passive process, for example following a drop in levels of DNMTs or in a replication-dependent manner (SINGER *et al.* 1979). It was, however, recognised early that certain waves of demethylation occurred during development and that the speed of demethylation of the male pronucleus in pre-implantation embryo is incompatible with a passive process (MORGAN *et al.* 2005). A mechanism of active demethylation must therefore exist. Early attempts to identify an enzyme capable of demethylating DNA proved futile, particularly in light of the fact that simple removal of the methyl group, requiring the breakage of a C-C bond is too energetically unfavourable. Nevertheless, the discovery of histone demethylases (LIN *et al.* 2008), which break a C-N bond and were also initially thought not to exist kept open the possibility of finding enzymes capable of breaking C-C bonds.

While the question of whether such an enzyme exists has yet to be answered, recently a number of cellular processes that seem to be involved in active DNA demethylation by alternative mechanisms have come to light. Although the precise enzymes are not known in full for every reaction, certain key players have been identified. The existence of both active and passive methods for demethylation is now generally accepted.

1.2.5.2 Passive demethylation

Evidence for passive demethylation has long been accumulating. An early model for the function of DNA methylation in differentiation took into account the likely role of passive demethylation. It was suggested that during the course of differentiation, as a transcription factor binds its recognition site, the CpGs within it would fail to be methylated, by virtue of the fact that the methylating

enzyme would lack access to said CpG (SINGER *et al.* 1979). Failure to methylate a CpG on a daughter strand following replication would lead to hemi-methylation in the daughter cell, which on the next round of cell division would yield one hemi-methylated cell and one unmethylated cell. This passive demethylation was therefore also termed replication-dependent demethylation (SINGER *et al.* 1979). The link between transcription factors and demethylation was strengthened when it was noted that Sp1 binding sites were associated with loss of methylation in CGIs (BRANDEIS *et al.* 1994; SILKE *et al.* 1995) and that NFκB was associated with loss of methylation in the immunoglobulin κ locus (KIRILLOV *et al.* 1996).

Experimental evidence for replication-dependent demethylation rapidly accumulated. Injection of methylated plasmid DNA into *Xenopus* embryos led to the CpGs close to the binding site of the transcription factors being selectively demethylated, in a process that was dependent on replication, but independent of transcription (MATSUO *et al.* 1998). Similar experiments carried out in mammalian cells confirmed that protein binding (in this case EBNA1 and LacI) to their recognition sites resulted in demethylation of these sites, independent of transcription (LIN *et al.* 2000) (LIN and HSIEH 2001). At the time, the proposed hypothesis was one of replication-dependent first strand demethylation followed by active demethylation by an as yet unidentified enzyme.

It was later shown that a transgene integrated into mouse ES cells resulted in its demethylation at the binding site for ETS transcription factors. When the DNA sequence was mutated in the vicinity of the CpGs to abolish transcription factor binding, demethylation no longer occurred (LORIOT *et al.* 2008). It was also shown that transient depletion of DNMT1, but not DNMT3A or 3B, led to stable hypomethylation of a transgene (LORIOT *et al.* 2006), suggesting another mechanism for passive demethylation. Replication dependent passive hypomethylation has also been shown to occur globally and at specific loci during the process of erythroid differentiation, in a manner correlated with speed of cell division (SHEARSTONE *et al.* 2011).

While there is compelling evidence for the existence of passive demethylation pathway(s), it remains unclear whether this has any functional consequences or is merely a bystander effect of other key processes such as transcription factor binding for lineage specification.

1.2.5.3 Active demethylation

As well as the speed of demethylation in pre-implantation embryos, the necessary existence of an active demethylating pathway came from the demonstration of demethylation in particular target genes in cells which were not undergoing replication. For example, the *ps2* gene (a target of ER-a) undergoes cyclical methylation/demethylation in non-dividing breast cancer cells, associated with the recruitment of DNMT3A and 3B, p68, TDG and BER (Base Excision Repair) proteins (METIVIER *et al.* 2008). Similarly, PTH (parathyroid hormone) induces active demethylation of *CYP27B1* which seems to be mediated by MBD4 following phosphorylation by PTH-activated protein kinase C (KIM *et al.* 2009). More recently, global DNA methylation studies of the differentiation of monocytes into macrophages (an event not requiring cell division or DNA replication) showed that numerous promoters lost methylation during the differentiation process and that this corresponded to the appearance of H3K4 methylation (KLUG *et al.* 2010). Other examples involving active DNA demethylation include the activation of memory promoting genes in rats (MILLER and SWEATT 2007), brain-specific genes following neuronal stimulation in mice (GUO *et al.* 2011) and *IL2* and *IFN-γ* promoters in human CD8 cells following antigen re-encounter (BRUNIQUEL and SCHWARTZ 2003). In all of these cases, the lack of cell division invokes an active demethylation mechanism.

The mechanisms for this active demethylation are not entirely clear and proposed hypotheses fall into a variety of pathways. Essentially these involve removal of the 5-methyl group, the 5mC base and the 5mC nucleotide. While in theory the simplest way of demethylating a C would be to remove its 5m group, in practice this is an energetically highly unfavourable reaction which has never been shown to occur in plants or animals. A number of processes can, however, bypass this reaction to remove the 5mC base or nucleotide.

1. Base Excision Repair pathway (BER)

(a) Direct excision of 5mC (unmodified base)

This is a mechanism for active demethylation which is well described in plants, however there are no known orthologs for the DME/ROS1 family of 5mC DNA glycosylases in mammals (CHEN and RIGGS 2011). In this pathway, DNA glycosylases create an abasic (AP) site by cleavage of the glycosidic bond between 5mC and deoxyribose. An AP lyase and an AP endonuclease (sometimes part of the same enzyme, as in some of the glycosylases) then cleave the DNA backbone and remove the deoxyribose, leaving a single nucleotide gap. Candidate glycosylases have been proposed to be TDG (in chicken) and MBD4, although in both cases their glycosylase activity is 30-40x weaker than their T:G mismatch activity (ZHU *et al.* 2000a; ZHU *et al.* 2000b). The nucleotide gap would then be filled again by an unmethylated cytosine by an unknown DNA polymerase and ligase.

(b) Deamination of 5mC followed by BER

The mechanism proposed involves active deamination of 5mC to thymine by hydrolysis, leaving a mismatched base T:G, which would be repaired by BER. This is akin to the spontaneous deamination which 5mC undergoes and the natural repair process that follows. Early data implicated the cytidine deaminases AID/APOBEC in this process, but the deamination only occurred on single stranded DNA and their knockout did not result in increased methylation (MORGAN *et al.* 2004). More recently, however, AID has convincingly been shown to be involved in this process. Zebrafish embryos in which AID/APOBEC and MBD4 are over-expressed undergo active demethylation at several loci (RAI *et al.* 2008). In another experiment, AID-deficient mouse primordial germ cells failed to undergo active demethylation at critical loci ((POPP *et al.* 2010). The AID deficient embryos were, however, viable and fertile and the primordial germ cells still had low levels of methylation, implicating a parallel pathway. Nevertheless, evidence of deamination as the first step in an active demethylation pathway exists and following deamination, the BER pathway is implicated. It is only recently that one enzyme definitively involved in the BER-

mediated demethylation has been identified as TDG (CORTELLINO *et al.* 2011). The importance of the BER pathway in active demethylation is highlighted by the fact that its inhibition with PARP inhibitors impedes DNA demethylation in the male pronucleus (HAJKOVA *et al.* 2010).

2. Nucleotide excision repair (NER) and hydrolysis

The hydrolysis necessary for this pathway is energetically unfavourable and initial *in vitro* experiments implicating GADD45a as a mediator of NER-coupled demethylation could not be confirmed. Nonetheless, GADD45a has been shown to be recruited to rDNA by TAF12, from where it recruits components of the NER pathway to promote dynamic demethylation and consequent gene activation (SCHMITZ *et al.* 2009). GADD45a promotes active demethylation of particular promoters as well as global hypomethylation when transfected (BARRETO *et al.* 2007). While not all experiments with GADD45a have confirmed these findings (JIN *et al.* 2008) there is also evidence that GADD45a plays a role in active demethylation in stem cell differentiation (ZHANG *et al.* 2011).

3. Oxidative demethylation

The discovery of the hydroxymethylation of 5mC to 5hmC by the TET enzymes has opened up an entirely new avenue in the search for active DNA demethylation. 5hmC may play a role in both active and passive demethylation. Its role in passive demethylation stems from the fact that DNMTs do not recognise it as a substrate (VALINLUCK and SOWERS 2007a) and therefore would not methylate 5hmC, leading to lack of remethylation in daughter cells. A role in active demethylation has been proposed whereby 5hmC is deaminated to 5hmU by AID/APOBEC and then excised by TDG (GONG and ZHU 2011). Furthermore, immunohistochemistry studies have shown that in the male pronucleus, 5hmC appears at the same stage as 5mC disappears, a finding that is abrogated by the knock-down of TET3 (INOUE and ZHANG 2011; IQBAL *et al.* 2011; WOSSIDLO *et al.* 2011).

1.3 Functions of DNA methylation in normal cells

A number of functional processes are thought to be regulated by DNA methylation: gene imprinting, silencing of repetitive elements, maintenance of genome integrity, and gene expression.

1.3.1 Gene imprinting

Gene imprinting refers to the mechanism by which maternal or paternal alleles are silenced such that monoallelic expression occurs at these key genes, which are predicted to have a prevalence of about 1% (LUEDI *et al.* 2007). DNA methylation plays a well-documented role in this process, with the best known example being the *IGF2/H19* locus. This locus is organised to contain *IGF2*, upstream of which is *H19*, containing a differentially methylated region (DMR), and upstream of which is a single enhancer. On the paternal chromosome, the *H19* DMR is methylated, silencing *H19* and allowing the enhancer to act distally on *IGF2*, allowing its transcription. On the maternal allele, the DMR is unmethylated, *H19* is expressed and CTCF (a transcriptional repressor) binds *H19* and prevents the action of the enhancer distally, such that *IGF2* is not expressed (CHAO and D'AMORE 2008a; CHAO and D'AMORE 2008b). Lack of methylation at the paternal allele due to a DNA methyltransferase mutation leads to aberrant paternal *H19* expression (LI *et al.* 1993). How the differential methylation at these loci is established is unknown. Conditional knockout experiments have, however, shown that DNMT3B is dispensable for imprinting whereas DNMT3A is essential for both maternal and paternal imprinting, as is the co-factor DNMT3L (KANEDA *et al.* 2004). Furthermore, an oocyte-specific variant of DNMT1 is implicated in maintaining these differentially methylated regions once they have been established (DING and CHAILLET 2002). Different regions of the DNMT1 protein seem to be involved in maintaining imprinted and non-imprinted regions (BOROWCZYK *et al.* 2009).

1.3.2 Silencing of repetitive elements

A large proportion of the human genome is made up of repetitive elements that have inserted themselves during the course of evolution. Repetitive elements can be classified as high copy number repeats and moderate copy number

repeats. High copy number repeats include heterochromatic repeats (such as satellite DNA which makes up the centromeric sequences) and interspersed repeats. These include LINEs (long interspersed elements), SINEs (short interspersed elements, of which the commonest are *Alu* repeats), LTRs (Long terminal repeats), minisatellites/VNTR (variable number tandem repeats), and microsatellites. Moderate copy number repeats include retroviral sequences (e.g. human endogenous retroviruses, HERVs). Subtelomeric repeats are another class of repeats and can be common to all chromosomes or chromosome specific. The abundance of repetitive elements in our genome has been suggested to be linked to novel transcriptional controls, especially as these are enriched for transcription factor binding sequences (POLAK and DOMANY 2006). Overall it has been estimated that SINEs, LINEs and HERVs altogether make up >45% of the human genome (LANDER *et al.* 2001). There is an average of 10 *Alu*s per gene and 1 LINE per gene (YODER *et al.* 1997). While some of these transposable elements lack the capacity to be transcribed due to mutations in their promoters, others such as LINEs and HERVs do contain intact sequences for transcription, which, were it to occur at numerous loci, would represent a considerable risk in terms of transcriptional noise and in particular could have antisense effects. Furthermore, repetitive elements may be involved in the formation of translocations (YODER *et al.* 1997). All of these reasons have led to the suggestion that DNA methylation has evolved as a manner of silencing these elements, both by silencing the promoters and by encouraging chromosome condensation which prevents translocations (WEBER *et al.* 2007; WEBER and SCHUBELER 2007). It was noted early after the discovery of 5mC that repetitive elements contained high levels of DNA methylation (GAMA-SOSA *et al.* 1983b). Restriction enzyme analysis followed by Southern blotting confirmed a high degree of methylation in major and minor satellite repeats (SANFORD *et al.* 1984). Bisulfite sequencing has confirmed these findings in LINES (CHALITCHAGORN *et al.* 2004) and *Alu*s are highly enriched for methylated CpGs in the mammalian genome (ROLLINS *et al.* 2006). Furthermore, insertion of retroviral sequences during mouse embryogenesis leads to their rapid inactivation by DNA methylation (JAHNER *et al.* 1982), and in mouse embryos lacking DNMT1, retrotransposons become reactivated 50-100 fold (WALSH *et al.* 1998).

1.3.3 Maintenance of genome integrity

In mice and human cell lines, inactivation of DNMTs leads to chromosomal abnormalities and ploidy changes, linking DNA methylation to the maintenance of genome integrity (DODGE *et al.* 2005). Furthermore, in human malignancies, hypomethylation precedes abnormal ploidy, suggesting that in the context of DNA hypomethylation, abnormal recombination events can occur more easily (EDEN *et al.* 2003). For example, a 10% drop in DNMT1 levels in mice leads to aggressive T-cell lymphomas with trisomy 15 (GAUDET *et al.* 2003). Perhaps the most compelling evidence for the role of DNA methylation in maintaining genome integrity is the ICF syndrome (immunodeficiency, centromere instability and facial anomalies). This results from mutations in DNMT3B (XU *et al.* 1999) where loss of methylation at pericentromeric satellites is associated with decondensation of heterochromatin and chromosomal fusions, confirming the role that DNA methylation plays in the correct maintenance of chromosomal structure (SUMNER *et al.* 1998).

1.3.4 Regulation of gene expression

Although this is an area of intense study, the role of DNA methylation in the regulation of gene expression is far from categorically determined. Some examples are relatively uncomplicated. The maspin gene (serpin B5) promoter contains a low density of CpGs. In cells where these are unmethylated, maspin is expressed, while in cells where they are methylated, no expression is seen (ECKHARDT *et al.* 2006a). Likewise, the promoters of genes on the inactive X chromosome (Xi) are methylated and the genes not expressed, although the methylation is not the initial stimulus, but rather a means of stabilising the repression (CEDAR and BERGMAN 2009a; CEDAR and BERGMAN 2009b). This simple rule does not, however, hold true for all genes, and for the ones which do follow it, it is not clear whether methylation is the mechanism by which the genes are silenced, or merely a mark by another signal which itself is responsible for controlling gene expression.

As detailed above, CGIs are almost always unmethylated in all cell types regardless of expression status (ECKHARDT *et al.* 2006a; RAUCH *et al.* 2009; SHEN *et al.* 2007)). A typical example would be alpha globin, which is only expressed in erythroid cells despite its CGI being unmethylated in all cell types. A small proportion (4%) of CGIs are normally methylated and the genes with which they are associated are silenced (SHEN *et al.* 2007). Clearly, therefore, lack of methylation is not sufficient to lead directly to gene expression. Conversely, some genes which are expressed do have a methylated CGI promoter (RAUCH *et al.* 2009). Several attempts have been made to reconcile these findings to the model linking promoter methylation with gene expression. It has been suggested ((RAUCH *et al.* 2009) (MEISSNER *et al.* 2008; WEBER *et al.* 2005) that genes with "strong" CGIs (high CpG density) and genes with "weak" CGIs (low CpG density) are regulated differently. Strong CGIs are nearly always unmethylated, and show no correlation between methylation and expression levels. They tend to correspond to either housekeeping genes or key developmental genes, in which case they are marked by bivalent chromatin (MEISSNER *et al.* 2008). Weak ones, however, are targets for *de novo* methylation in somatic cells, and tend to correspond to germ-line specific genes. These genes do show a correlation between methylation and level of expression (WEBER *et al.* 2007). While these findings were based on MeDIP, a technique that has received a lot of criticism since its introduction, the main thrust of the results has been upheld when this question has been examined with other methods. For example, when using targeted bisulfite sequencing with padlock molecular probes, a negative correlation between DNA methylation at TSS and gene expression was noted and "weak CGIs" were shown to be targets for methylation while "strong CGIS" were nearly all unmethylated in fibroblast and iPS cells (DENG *et al.* 2009). These results were substantiated in both fibroblast and iPS cells as well when examined by whole genome bisulfite sequencing (LAURENT *et al.* 2010). In another study, however, while there was a correlation between CpG methylation and gene expression in fibroblasts, this did not hold true in ES cells, where only CHG/CHH methylation and gene expression were correlated (LISTER *et al.* 2009).

While CGIs are generally unmethylated, CpGs in gene bodies are heavily methylated. Not only does this not prevent gene expression, but it seems to be positively correlated with it. Using MIRA (Methylated CpG-Island Recovery Assay), a technique where methylated DNA is enriched by use of a complex of methyl-binding proteins, it has recently been confirmed that genes with higher expression levels show higher gene body methylation (RAUCH *et al.* 2009), a finding previously noted in *A thaliana (ZILBERMAN et al. 2007)*. This was also demonstrated for genes on the active X chromosome (Xa), where CGIs are unmethylated and gene bodies heavily methylated. The corresponding genes on Xi have the opposite pattern (HELLMAN and CHESS 2007). *In vitro* models of intragenic methylation have found that this is associated with depletion of active chromatin marks and loss of PolII from the methylated region (LORINCZ *et al.* 2004) Nevertheless, the flurry of papers investigating DNA methylation at individual CpG resolution has confirmed gene body methylation in highly expressed genes (BALL *et al.* 2009; DENG *et al.* 2009) (LAURENT *et al.* 2010; LISTER *et al.* 2009) and this may serve to suppress illegitimate transcription from alternate promoters (MAUNAKEA *et al.*).

Much attention has focused on areas that are differentially methylated between tissues, so-called t-DMRs. A large proportion of these occur outside promoters (ECKHARDT *et al.* 2006a; SHIOTA 2004; SONG *et al.* 2005). These t-DMRs are assumed to play a role in the differential expression of genes. They can be found in 5' UTRs, exons or introns. One feature over-represented in t-DMRs is evolutionarily conserved regions. These are often far (up to 100kb) from any protein-coding gene and are thought to play a long-range interaction role in the control of gene expression. Further evidence that methylation is functional in this process is, however, lacking. T-DMRs can be used in hierarchical clustering to differentiate tissue types and are conserved between mouse and human (IRIZARRY *et al.* 2009).

Interestingly, some genes with a methylated promoter but which are still expressed are in fact being expressed from an alternative promoter (RAUCH *et al.* 2009). This suggests that methylation of the main promoter is indeed a

mechanism for controlling gene expression. Likewise, it is possible that methylation at 3' UTRs serves to suppress antisense transcription. DNA methylation may also be involved in the maintenance rather than the initiation of gene silencing (FELDMAN *et al.* 2006). This suggestion comes from the study of the Oct4 promoter (which is CpG-poor), where DNA methylation is not necessary for the initial silencing of the gene, but is required to prevent its re-expression at inappropriate stages.

DNA methylation has also been suggested to play a role in the control of gene expression that is necessary for maintenance of pluripotency. As cells differentiate from mouse ES cells to neuronal progenitors, hundreds of promoters become hypermethylated, particularly at pluripotency genes. The methylation pattern upon further differentiation changes relatively little beyond this point and only a very small proportion of tissue-specific genes are methylated in ES cells and lose methylation and gain expression with differentiation (MOHN *et al.* 2008). It is clear, however, that as cells differentiate, DNA methylation changes are dynamic, with both gains and losses of methylation at distinct sites, which can in part be predicted from chromatin context such as the presence of H3K4me3 and absence of H3K9me3 predicting the appearance of unmethylated DNA (MEISSNER *et al.* 2008).

1.3.5 Methylation and structure

The pattern of DNA methylation also seems in part influenced by the chromosomal context of DNA. For example, several studies have noted that there is a 10 base-pair periodicity to DNA methylation, from organisms as diverse as *Arabidopsis* and humans (CHODAVARAPU *et al.* 2010). This may be explained by the structure of the enzymes involved in *de novo* methylation. DNMT3A and DNMT3L are found in a tetrameric complex which contains two active sites separated by about one DNA helical turn (JIA *et al.* 2007). Furthermore, nucleosomal DNA is more highly methylated than flanking DNA (CHODAVARAPU *et al.* 2010) and DNMT3A and 3B, but not DNMT1, are found tethered to nucleosomes, and in particular those containing methylated SINE and

LINE elements (JEONG *et al.* 2009a). Thus, the configuration and location of DNA methyltransferases may have an effect on the patterns of DNA methylation.

1.4 Cross-talk between DNA methylation and histone modifications

DNA methylation and histone modifications interact in various ways. The term cross-talk is quite appropriate as there is evidence both for DNA methylation guiding histone modifications and the reverse. In fact, they are often found as part of a mutually-reinforcing loop.

1.4.1 When Histone modifications orchestrate

An example of histone modifications being the initial event in silencing occurs at pericentromeric heterochromatin. In mouse ES cells, the histone methyl transferases SUV39H1 and SUV39H2 first set down the H3K9me3 mark at these loci, which then recruits DNMT3B, possibly through HP1, and subsequently leads to DNA methylation (LEHNERTZ *et al.* 2003b). The repressive environment is then further strengthened by the recruitment of MeCP2 and HDACs to these methylated loci, with histone deacetylation reinforcing the repression signaled by DNA methylation (LEHNERTZ *et al.* 2003a). HP1 has been shown to associate with DNMT1 and stimulate its methyltransferase activity *in vitro* and *in vivo* (SMALLWOOD *et al.* 2007). This is an example of a reinforcing loop where a histone modification has led to DNA methylation, which itself has led to other histone modifications.

Another example of histone modifications directing DNA methylation involves PRMT5, an arginine methyltransferase which symmetrically dimethylates H4R3. This mark allows recruitment of DNMT3A and *de novo* methylation of key CpGs, which leads to silencing of gamma-globin (ZHAO *et al.* 2009)).

Genome-wide it has been shown that deficiency of G9A, which sets down the H3K9me3 mark at euchromatic loci (TACHIBANA *et al.* 2002) causes a reduction in DNA methylation at specific loci (IKEGAMI *et al.* 2007). This suggests that

H3K9me3 is involved in directing DNA methylation at specific targets and that this is not a general phenomenon. In these examples a histone mark directly leads to DNA methylation.

1.4.2 When DNA methylation directs

Because with every round of DNA replication new histones must be laid down, it is not clear how histone modifications can be passed on to the daughter cell. DNA methylation has been proposed as a mechanism by which this information may be faithfully replicated in daughter cells. For example, methylated DNA will recruit MeCP2 or MBD2, which themselves can recruit HDACs and thereby copy the repressive nucleosomal environment which characterises regions of methylated DNA (CEDAR and BERGMAN 2009a). However this simple hypothesis does not take into account the very different chromosomal environments between a methylated CGI, which would indeed be in a repressive environment, and a densely methylated gene body which on the contrary would be in a region highly permissive for DNA transcription.

Studies of DNMT knockout mice have shown that there is evidence of DNA methylation directing histone modifications. DNMT1 knockout mice globally lose H3K9me2 and H3K9me3 (ESPADA *et al.* 2004). DNMT3 negative mouse ES cells have higher levels of acetylated H3K9, H4K5, and H4K16 but less H3K9me2. In these cells there was no decompaction of chromatin but an altered binding of linker histones was noted (GILBERT *et al.* 2007).

While there is evidence for both histone modifications directing DNA methylation and *vice versa*, it is not always clear how two related findings are mechanistically connected. One such example is the anticorrelation between DNA methylation and the histone variant H2A.Z. This variant is found at transcriptionally active genes. It was initially shown in plants (ZILBERMAN *et al.* 2008) and then confirmed in mammals (EDWARDS *et al.* 2010), that the location of H2A.Z and DNA methylation are mutually exclusive. It has been proposed that incorporation of H2A.Z, whose function may be to facilitate transcription by destabilizing nucleosomes, may prevent the recruitment of DNMTs (MARQUES *et*

al. 2010). Alternatively, methylated DNA may somehow prevent incorporation of H2A.Z, thereby improving nucleosome stability and reducing access of DNA to transcription factors. In a mouse model for B-cell lymphoma, as genes gained H2A.Z at their promoters, DNA methylation disappeared at the same loci. The inverse was not true, however, as loss of H2A.Z from TSS was not associated with a gain in DNA methylation, which the authors presume to be due to protection from methylation by other factors (CONERLY *et al.* 2010).

An absolute correlation, be it positive or negative, between two events, does not necessarily imply that there must be a direct link between them. In an interesting study published recently in epithelial cells, there was colocalisation of intragenic DNA methylation and H3K36me3, a mark of transcription elongation (HAHN *et al.* 2011). In the 5' regions of genes, H3Ac was seen without DNA methylation, gene bodies had both H3K36me and DNA methylation, and the 3' regions were characterised by a loss of both H3K36me3 and DNA methylation. Knockdown of DNMT1 and DNMT3B, however, reduced DNA methylation without affecting H3K36me3 or transcription. Conversely, siRNA knockdown of SETD2, a major H3K36me3 methyltransferase, showed only a few genes which became repressed, although most showed no change in DNA methylation or expression levels. This illustrates that despite there being a clear correlation between DNA methylation and H3K36me3, reducing either does not affect the other, refuting a direct mechanistic link between the two observations.

Occasionally, co-incident findings are related in an unexpected manner. For example it has been shown that KDM1, a histone H3K4 demethylase, is necessary for maintenance of DNA methylation, as reducing KDM1 levels leads to a gradual loss of DNA methylation (WANG *et al.* 2009). This was not found to result from a link between H3K4me3 and DNA methylation, as levels of H3K4me3 remained unchanged in KDM1 deficient cells. Rather, the authors determined that DNMT1 is a non-histone target of KDM1. DNMT1 is methylated by SET 7/9, thereby reducing its stability and leading to loss of DNA methylation. The normal presence of KDM1 keeps DNMT1 unmethylated and increases its stability, explaining the importance of KDM1 in maintaining DNA methylation levels.

A caveat of studies that focus on a small region of the genome is that correlations between two independent events may not hold up when examined on a genome-wide scale. A recent survey of published ChIP-Seq studies in combination with genome-wide DNA methylation studies in mice (84% of all CpGs covered) has reported the following associations between DNA methylation and histone modifications: there was no correlation with methylation at H3K36, H3K27, H3K79, or H3K9 and DNA methylation. H3K4 di- and tri- methylation were strongly anticorrelated with DNA methylation, as was H2A.Z (EDWARDS *et al.* 2010). The authors propose that DNA methylation is the default state in the genome and that unmethylated regions arise from exclusion of DNMTs by very high density CpGs or certain histone variants (H3K4me2, H3K4me3, H2A.Z). As whole-genome bisulfite sequencing studies begin to accumulate, some of these associations will be clarified.

The links between histone modifications and DNA methylation are not complete without mentioning MBD proteins. For example MBD1, which binds methylated DNA, has been suggested to recruit SUV39H1 to regions of methylated DNA and thereby participate in transcriptional repression (FUJITA *et al.* 2003).

1.5 Epigenetic alterations in malignancy

1.5.1 Global DNA hypomethylation

Initial studies of DNA methylation in malignancy noted a lower overall 5mC content by HPLC and by sensitivity to restriction enzyme cutting in a number of different tumours (GAMA-SOSA *et al.* 1983a). It quickly became evident that accurate quantification of methylation loss in malignancy could only be achieved by comparing the malignant tissue with its appropriate normal counterpart, due to the variability of methylation between tissue types (e.g. by HPLC, cerebellar cells have 17% more 5mC than cardiac cells (EHRLICH 2002)). Comparison of total 5mC content by HPLC in colon cancer, benign adenomas and normal adjacent tissues showed a 10% reduction in both the benign and malignant tumour cells (FEINBERG *et al.* 1988). An exhaustive list of malignancies in which

global hypomethylation has been reported would be impractical to report, but would include for example prostate cancer (BEDFORD and VAN HELDEN 1987), CLL (WAHLFORS *et al.* 1992), cervical cancer (KIM *et al.* 1994), and colorectal cancer (HERNANDEZ-BLAZQUEZ *et al.* 2000). It has been suggested that methylation anomalies accumulate with progression of the disease (whether as a cause or consequence) and overall 5mC content decreases from 3.7% in benign tumours, to 3.4% in primary malignancies and 3.2% in secondary malignancies (EHRLICH 2002). This universal phenomenon of global hypomethylation is variable, however, as some cancers such as those of the bladder seem much more affected than others, including renal malignancy or multiple myeloma (ROSS *et al.* 2010).

1.5.2 Methylation at repetitive elements

The global hypomethylation noted in malignancy was initially assumed to be due to the hypomethylation that is frequently found at repetitive elements (EHRLICH 2009), although this idea is undergoing some revision in the current literature (HANSEN *et al.* 2011). Nonetheless, it cannot be disputed that repetitive elements, which make up a large part of the genome and which are highly methylated in the normal setting, are certainly hypomethylated in the vast majority of malignancies.

Hypomethylation at repetitive elements is frequent but varies between types of repeats and between types of malignancies. For example, in bladder cancer, the reduction in DNA methylation, as assessed by pyrosequencing, was 13% in alpha satellite repeats, 63.5% in the NBL-2 tandem repeat, 12.5% in LINE1 and 5.4% in Alu Yb8 (CHOI *et al.* 2009). On the other hand, methylation at the D4Z4 tandem repeat increased by 6.9%. In hepatocellular carcinoma, hypomethylation of Sat2 occurs earlier than LINE1 or *Alu* sequences (LEE *et al.* 2009). Although a common feature, the significance of hypomethylation at repetitive elements is variable as in some cancers it arises early, while in others it arises late. In some malignancies the level of hypomethylation in repetitive elements carries prognostic significance (OGINO *et al.* 2008; SAITO *et al.* 2010). In some cases the

extent of the hypomethylation has been correlated with the level of genomic instability although a causal effect is not clear (DASKALOS *et al.* 2009). Because methylation of repeats at centromeres is purported to contribute to genome stability by providing a structural condensing role, it follows that hypomethylation of these sequences could be associated with loss of genome stability. Indeed, correlations between hypomethylation of centromeric repeats and gain in chromosomal translocations have been reported (DASKALOS *et al.* 2009) (GAUDET *et al.* 2003), however it does not follow that there is a causal link between these two observations.

It is more difficult to compare levels of methylation between tumour types when different techniques have been used, but certainly hypomethylation at LINE elements has been reported in colorectal cancer (SUTER *et al.* 2004), ovarian cancer, (DAMMANN *et al.* 2010), prostate cancer (YEGNASUBRAMANIAN *et al.* 2008) and NSCLC (DASKALOS *et al.* 2009) to name a few.

Hypomethylation at repeats is linked to increased transcription from these sequences. For example, different classes of HERVs were over-expressed 70-3000 fold in testicular cancer (GIMENEZ *et al.* 2010).

1.5.3 Localised hypermethylation

The third panel in the triptych of DNA methylation changes in malignancy, besides global hypomethylation and hypomethylation of repetitive elements, is that of hypermethylation of particular gene promoters. The parallel finding of hyper- and hypomethylation in malignancy, albeit at quite distinct sites, has always been an intriguing puzzle and has fuelled the hypothesis that different mechanisms must be at play to explain these disparate findings. As there is no correlation between hypermethylation of promoters and hypomethylation of repeats, separate mechanisms must be invoked (LEE *et al.* 2009). Numerous studies have looked at which genes are affected by this hypermethylation. Early studies were no doubt plagued by a certain amount of ascertainment bias, and genes commonly affected by promoter hypermethylation were found to be tumour suppressor genes thought to be involved in the development of these

malignancies. As such, conclusions were drawn linking novel hypermethylation in malignancy at tumour suppressor genes and development or progression of the tumours, whereas more unbiased methylation analysis techniques today suggest less of a link between the two events as will be discussed below.

As methylation studies have become genome-wide, particular trends have been repeatedly noted. The genes affected by promoter hypermethylation fall into common categories such as DNA repair, apoptosis, cell cycle control and metastasis (BOULTWOOD and WAINSCOAT 2007). Of note, however, within and across these categories, quite a limited set of genes appears to be affected by hypermethylation in different cancer types. For example, MDS and AML cases are associated with hypermethylation of *p15* and *E-cadherin*, while haematological malignancies more generally are associated with methylation at *DAPK1* and *RARB2* (BOULTWOOD and WAINSCOAT 2007). In an assessment of target genes hypermethylated in AML, Melki *et al.* (MELKI *et al.* 1999) found that 95% of AML cases had at least one of a set of genes hypermethylated, while 75% had at least two of these hypermethylated, suggesting commonality in the genes targeted by hypermethylation in a particular malignancy. Furthermore, in a different study that identified 33 hypermethylated loci in AML, none of these loci were affected in 6 solid tumours (RUSH *et al.* 2001). Lung cancer is particularly associated with hypermethylation of *HOX* genes (RAUCH *et al.* 2007). It therefore seems that the genes targeted for hypermethylation are not random, analogous to the tissue-specific methylation patterns in normal cells. Overall, 5-10% of normally unmethylated CGIs become methylated in most malignancies, but the actual pattern will vary (BIRD 2002).

Some of the genes affected by hypermethylation are so crucial to the development of malignancy that it is very tempting to assume that there is a causal link between the events of hypermethylation and disease progression, in particular because of the link between hypermethylation and repression of transcription. For example, the gene encoding hMLH1, a DNA mismatch repair enzyme, is hypermethylated in colorectal cancers, and this is associated with both loss of expression of the gene and microsatellite instability (HERMAN *et al.*

1998). In these experiments, cells which did not have microsatellite instability were far less likely to have a hypermethylated *hMLH1*. Furthermore, treatment of cell lines with a demethylating agent not only caused re-expression of hMLH1, but a recovery in mismatch repair ability. Experiments such as these are very suggestive of the causal link between DNA hypermethylation and development and/or progression of malignant phenotype.

How is hypermethylation targeted to specific regions? One recent study has considered the possibility that particular transcription factors interact with DNMTs and target them to their binding sites (HERVOUET *et al.* 2009). For example, c-myc interacts with DNMT3A and recruits this *de novo* methyltransferase to the promoters of c-myc regulated genes, such as *CDKN2a*, *TIMP2* and *CCND1*, thereby promoting the methylation of CpGs in their promoters and reducing their expression. Abrogation of c-myc was associated with lack of methylation at these sites and increased gene expression. The authors then identified a number of other transcription factors that would be capable of binding DNMT3A or DNMT3B and perform similar targeting functions.

1.5.4 Away from the islands and towards the shores, and back again

It is evident that the findings reported on DNA methylation in malignancy will reflect the limitations of the techniques used to study this phenomenon. Unsurprisingly, the first reported findings included global levels of methylation, methylation at repetitive sequences and at CGIs, easily assayed by Southerns. More genome-wide studies such as MeDIP were also concentrated on CpG dense regions. CHARM (comprehensive high throughput arrays for relative methylation), a method based on fractionation of the genome with an enzyme which cuts methylated DNA, is unbiased and non-promoter driven and has been extensively validated by other methods (IRIZARRY *et al.* 2008a). A key paper using this technique showed that most tissue-specific t-DMRs were found in CGI "shores", i.e. within 2kb of CGIs. Looking at colorectal cancer, about 2700 c-DMRs (cancer-specific DMRs) were identified, roughly equally divided between regions that gained and lost methylation. These changes in methylation were

linked with changes of gene expression in nearby genes. This was not straightforwardly correlated but the greatest correlation was between changes in methylation from "none" to "some" and loss of expression. Interestingly, there was a greater overlap in t-DMRs and c-DMRs than would be predicted by chance, suggesting a link between epigenetic changes involved in tissue differentiation and those involved in malignant transformation. This study also highlighted the importance of not concentrating on CGIs for studies of DNA methylation (IRIZARRY *et al.* 2009).

The finding that hypermethylation occurs in CGIs and hypomethylation in CGI shores was recently challenged when whole-genome bisulfite sequencing was used to show that both of these phenomena can be explained by a shift in methylation boundary in malignancy. As such, if methylation moves away from a shore and into an island, the shore will look relatively hypomethylated and the island relatively hypermethylated compared to a normal cell, when in fact the level of methylation is unchanged but has shifted position. Loss of methylation stability was used to explain these findings (HANSEN *et al.* 2011).

1.5.5 The epigenetic machinery in cancer

1.5.5.1 DNA methyltransferases

The first study of DNMT levels in malignancy was in colorectal cancer where increased levels of DNMT1 were observed (EL-DEIRY *et al.* 1991). This reached a several hundredfold expansion in the malignant cells as measured by RQ-PCR, while pre-malignant cells had DNMT1 levels elevated to a lesser degree. Further studies that better controlled for rates of proliferation and used an RNAse protection assay found these increases to be in the order of magnitude of 2-fold (LEE *et al.* 1996). Interestingly, when colon cancer cells were examined with immunohistochemistry for DNMT1 expression, a striking level of heterogeneity was noted, with some cells grossly over-expressing DNMT1 and other hardly expressing any (DE MARZO *et al.* 1999). This highlights the importance of remembering that most techniques usually study a population of cells that may not be homogeneous, and methylation values represent an average of all the cells in the population. Elevated levels of DNMT1 have been reported in other

malignancies such as AML (4.4 fold) (MELKI *et al.* 1998), and CLL (2-4 fold for DNMT1 and DNMT3A) (MELKI and CLARK 2002). It must be highlighted, however, that fold expression increase is dependent on which gene is used for normalisation. While DNMT over-expression could explain the hypermethylated CGIs, it is at odds with the finding of global and specific hypomethylation, and there is no correlation between DNMT levels and the amount of CGI hypermethylation (EADS *et al.* 1999). DNMT3A and 3B were also shown to be elevated 2-8 fold in bladder, colon, kidney and pancreatic tumours (ROBERTSON *et al.* 1999) and DNMT3B in endometrial cancers (JIN *et al.* 2005).

Abnormalities of the methylation machinery in malignancy go beyond mere enzyme levels. For example, a DNMT3B splice variant, DNMT3B4, is elevated in hepatocellular carcinoma. This has no DNMT catalytic activity but may act in a dominant negative fashion to prevent DNMT3B from methylating pericentromeric satellites, resulting in hypomethylation and genomic instability (KANAI and HIROHASHI 2007).

More commonly reported are mutations in key methylation machinery factors. For example, DNMT3A mutations are seen in 22% of AML, which is associated with adverse outcome (YAMASHITA *et al.* 2010), (LEY *et al.* 2010),(YAN *et al.* 2011). The effects of the mutations on overall and specific DNA methylation levels, and the correlations between changes in methylation and changes in gene expression are not necessarily clear, leading to the suggestion that the effects of DNMT3A mutations may be independent of their effects on methylation, or that the interaction of mutations in other commonly affected genes such as *FLT3* or *NPM1* may mask a direct effect (SHAH and LICHT 2011). It is also possible that the techniques used to study methylation in these experiments were not most appropriate, as some of these were immuno-precipitation based. Nevertheless, the consistent finding that DNMT3A mutations are a marker for poor prognosis is interesting and if it is confirmed that this is independent of other known prognostic markers could lead the way to this being used as a new biological marker.

In other circumstances, enzyme levels and structures are normal, but their recruitment is abnormal, such as PML/RARα recruiting DNMT3A to the *RARB2* promoter and aberrantly silencing it (DI CROCE *et al.* 2002). Conversely, DNMT1 may be excluded from its normal recruitment pattern such as in glioma cells where abnormal phosphorylation of DNMT1 by Akt or protein kinase C prevented it from interacting with UHRF1 and PCNA, which are normally responsible for targeting DNMT1 to the replication fork, leading to hypomethylation (HERVOUET *et al.* 2010a).

Finally, it has been shown that damaged DNA bases (such as 5-chlorocytosine, which can result from inflammation) can mimic 5mC and cause hemi-methylation of the opposite strand by DNMT1 (VALINLUCK and SOWERS 2007b). In this example, the enzyme is normal but it is the substrate that is abnormal. It remains to be seen whether this is an important effect *in vivo*.

1.5.5.2 Histone proteins

Histones may also be abnormal in malignancy. H4K16ac and H4K20me3 marks were first noted to be lost in lymphoid and myeloid cell lines compared to their normal counterparts, and this was found to be most marked at repetitive sequences which are normally DNA methylated (FRAGA *et al.* 2005b). These findings were confirmed in primary tumours. In lung cancer, H4K12 and H4K16 acetylation and H4K20me3 are globally depleted, while H4K5 and H4K8 are hyperacetylated (VAN DEN BROECK *et al.* 2008). Loss of imprinting at the *ARHI* gene locus has been reported in breast cancer, with abnormal expression of the gene product, and this is associated with loss of DNA methylation, increase of H3Ac and loss of H3K9me3 (FUJII *et al.* 2003). Chromatin remodelling is also mistargeted by translocations, for example *MLL* is a frequent translocation partner in AML. When *MLL* is involved in these translocations, it causes persistent activation of *Hoxa9*, due to abnormal deposition of H3K79me2, possibly through an interaction with DOT1 (MILNE *et al.* 2005).

1.5.5.3 MBD proteins

Other aspects of the epigenetic programme may also be perturbed in malignancy. For example, MBD protein levels are frequently altered. MeCP2 is reduced in hepatocellular carcinoma, and the levels correlate with disease progression (KANAI and HIROHASHI 2007). MeCP2 and MBD2 are significantly reduced in all types of NSCLC, however they are only minimally perturbed in AML (MULLER-TIDOW *et al.* 2001). The distribution of MBD proteins in malignancy has also been studied. In cancer cell lines it has been shown that most hypermethylated promoters are bound by MBDs, whereas these are absent from unmethylated promoters, except for MBD1, which can bind unmethylated DNA. The distribution of which MBDs bound which promoters seemed constant within a type of malignancy, suggesting a degree of specificity in how MBDs interact with methylated DNA in malignancy (LOPEZ-SERRA *et al.* 2006).

1.5.6 Are epigenetic mechanisms recapitulating stem cell epigenetics?

As stated above, specific gene promoters are often co-hypermethylated in malignancy. For example, in haematological malignancies, the promoters of *DAPK*, *p15, p16* and *E-cadherin* are recurrently found to be hypermethylated (GALM *et al.* 2006; SEELIGER *et al.* 2009). The term "CpG island methylator phenotype" has been used to describe the seemingly coordinated aberrant methylation in a specific subset of genes noted in malignancy. Intriguingly, the finding that many of the genes which undergo hypermethylation in malignancy are polycomb targets in normal stem cells, has highlighted the similarities between normal stem cells and cancer-initiating progenitor cells (JONES and BAYLIN 2007). In colorectal cancer, a set of 135 hypermethylated promoters was identified and found to be distributed non-randomly in the genome. Strikingly, most of the genes associated with these promoters were already repressed in normal cells despite the promoters being unmethylated, suggesting that methylation was not a mechanism for the initiation of repression. The authors suggested an instructive mechanism may target these genes for methylation (KESHET *et al.* 2006). Further evidence came from colorectal cancer, where genes hypermethylated in malignancy were specifically associated with the H3K27me3 mark in normal and malignant colon cells. Nevertheless, the link with DNA

methylation was only seen in the cancer cells as the genes which did not become methylated were not marked by H3K27me3 in normal or cancer cells (SCHLESINGER *et al.* 2007). In the normal situation, H3K27me3 is usually not associated with DNA methylation, despite being a repressive mark, and these two modifications are often anticorrelated (LYNCH *et al.* 2011).

The finding was independently reported that in colorectal, ovarian and breast cancer, hypermethylated genes were found to frequently overlap with genes marked by polycomb repression in ES cells (WIDSCHWENDTER *et al.* 2007). The authors suggest that immortalisation arises in an early cell where PRC2 is still targeting stem cell genes and suggest cross-talk between PRC2 and DNMTs to explain this link. EZH2, a component of PRC2, was shown to interact directly with the DNMTs and DNMTs were specifically recruited to EZH2 target genes, which they methylated (VIRE *et al.* 2006), lending credence to this hypothesis. This could explain the mechanistic link between the polycomb targets in ES cells and the hypermethylated targets in cancer cells. The model was further refined when it was noted that hypermethylated genes in adult cancer cells were unmethylated in ES and embryonal carcinoma (EC) cells but marked not only by polycomb but also by bivalent chromatin marks in ES cells (i.e. both the active H3K4me3 and the repressive H3K27me3 marks). EC cells additionally had H3K9me2 and H3K9me3 at the sites of genes hypermethylated in cancer. The suggestion is that once these repressive histone marks are added, the genes are targeted for further deep silencing by DNA methylation (OHM *et al.* 2007). In a detailed study of a gene hypermethylated in EC cells (*GATA4*) and bearing polycomb marks, it was shown that in EC cells, the locus is organised in a multi-conformational loop which is lost as the cells differentiate. In colon cancer cells, the promoter is hypermethylated, the gene not expressed and the loop is tightened. When DNA methylation or the H3K27me3 marks are absent, the loop is relaxed. The suggestion is that the repressive loops are involved in gene silencing (TIWARI *et al.* 2008). Importantly, however, a recent unbiased study of methylated and repressed genes in breast cancer has highlighted that while hypermethylated genes in breast cancer were enriched for bivalent marks in ES cells, the bivalent mark was absent from the majority of the genes affected by

hypermethylation. The authors therefore suggest that there is no mechanistic link via polycomb between the bivalent mark and hypermethylation in malignancy but rather that bivalently marked genes are targets for methylation by virtue of the fact that they are lineage specific (SPROUL *et al.* 2011). How lineage specificity could guide hypermethylation and what role this might play in the development or maintenance of malignancy remains a mystery.

It is also important to focus on the fact that the genes which are hypermethylated in malignancy are already repressed before the methylation mark appears, and the significance of this observation is unclear (CEDAR and BERGMAN 2009a). In the recent study of breast cancer cell lines and primary tumours mentioned above, genes with methylated promoters were less highly expressed than genes with unmethylated promoters (SPROUL *et al.* 2011). Interestingly, genes that were variably methylated across the cell lines were generally unmethylated in normal tissues and were enriched for genes involved in developmental processes and enriched for genes with bivalent chromatin marks in hES cells. More importantly, genes that were methylated and repressed in malignant cell lines were already repressed in normal cells, and this observation was consistent with different lineages of breast cancer cells. Further evidence that methylation is not the cause of repression is that treatment with demethylating agents did not lead to increased expression of these genes despite causing demethylation (SPROUL *et al.* 2011).

1.5.7 Towards a new outlook on methylation anomalies in malignancies

The advent of genome-wide methylation studies has further refined the epigenomic landscape in malignancy and the chromosomal distribution of methylation changes appears to be non-random. For example it was found in prostate cancer that DNA hypermethylation was not restricted to individual genes but spread over large regions (e.g. 4Mb) and that gene silencing spread as well, a concept dubbed LRES: long-range epigenetic silencing. This was associated with silencing of neighbouring unmethylated genes and seemed to result from chromatin remodelling. The regions of DNA methylation were associated with lack of H3K9ac, and enrichment of H3K9me2 and H3K27me3. It

must be said that not all LRESs had identical findings for DNA methylation, histone modifications and expression levels. What did appear from the analysis, however, is that changes in DNA methylation and/or histone modifications occurred in blocks of about 2Mb, suggesting a coordinated action (COOLEN *et al.* 2010).

Other studies have also noted the finding of methylation changes occurring in a coordinated fashion. In a colon cancer cell line, large domains of differential methylation spanned 20Mb and occurred in gene poor regions (WEBER *et al.* 2005). This was corroborated by a recent study in colon cancer, which identified large blocks across the genome where hypomethylation occurred in a coordinated fashion (HANSEN *et al.* 2011). This was not due to the presence of repetitive elements. The location of these blocks of hypomethylation corresponded to previously identified partially methylated domains which lose methylation with cell differentiation.

The same study of c-DMRs across cancer types revealed an interesting observation. While c-DMRs could accurately distinguish normal from cancer cells, the latter had quite varied methylation patterns at these loci within a tumour type but between individuals. This variability was not across the board in the malignant cells but was concentrated at c-DMRs. The location of c-DMRs corresponded to t-DMRs in normal tissue differentiation. The authors propose that this is due to a "relaxation" of epigenetic surveillance so that the increase in plasticity that results allows a stochastic variability in gene expression at these loci. In stem cells, this would allow different progeny to differentiate along alternative pathways depending on external stimuli, while in malignant cells it would allow them to adapt to selective pressures that develop during tumourigenesis. Indeed, expression of genes in these areas was found to be hypervariable in malignancy compared to normal differentiated cells (HANSEN *et al.* 2011). This provides an elegant link between mechanisms utilised by pluripotent stem cells and those used by malignant cells.

In the same way that normal embryonic stem cells seem to harbour both 'active' and 'repressive' histone marks at the same promoters, presumably to keep the cells 'poised' for rapid induction of certain programmes of transcription associated with either renewal or differentiation, the implication is that some 'cancer stem cells' may harbour similar chromatin signatures and could contribute to the renewal potential of tumours.

1.5.8 Clinical implications of DNA methylation

The first clinical implication of DNA methylation is its use as a biomarker for malignancy, in particular as the genes affected by hypermethylation are non-random. Numerous studies have evaluated this possibility. For example HeavyMethyl RQ-PCR analysis of 4 biomarkers in urine showed them to be able to discriminate prostate cancer patients from normal controls with a sensitivity of 86% and a specificity of 89% (ROUPRET *et al.* 2007). Gene methylation may also provide prognostic information, such as in diffuse large B-cell lymphoma patients, where promoter hypermethylation-induced silencing of the DNA repair gene O^6MGMT is a good prognostic marker for patients receiving chemotherapy with cyclophosphamide, as the chemotherapy-induced lesions could perhaps not be repaired, leading to apoptosis of the cells (ESTELLER *et al.* 2002). Conversely, the degree of overall hypomethylation in prostatic cancer is correlated with its metastatic potential (BEDFORD and VAN HELDEN 1987).

More important, however, is the role of demethylating agents in the treatment of cancer. These are nucleoside analogues which are incorporated into DNA and inhibit DNMT1, being active in S phase (EGGER *et al.* 2004). They have other mechanisms of action, such as forming adducts with DNMT1, which targets them for proteosomal degradation (PATEL *et al.* 2010). One such agent, 5-azacitidine, initially developed as a cytostatic agent for the treatment of AML, was later demonstrated to have demethylating activity when used at lower doses. In tissue culture, use of a combination of demethylating and HDAC inhibitors has a synergistic effect, as gene silencing usually also involves histone deacetylation (CAMERON *et al.* 1999). In acute myeloid leukaemia, a combination of 5-azacitidine and phenylbutyrate led to major responses, with the degree of

response correlating with the degree of hypomethylation at *p15* (GORE *et al.* 2006). Demethylating agents have shown most promise in high risk myelodysplastic syndrome (MDS), for which there was previously no treatment. With azaditidine, overall survival improved from 15 months to 24 months in a randomised controlled trial, and time to AML was delayed by 6 months (FENAUX *et al.* 2009).

Clearly the use of a demethylating agent will have an effect genome-wide whereas the targets of hypermethylation are quite discrete and as has been established, correct patterning of methylation would seem to be a prerequisite for adequate cell function. Nevertheless, the use of this very blunt tool does seem to target enough of the key genes to have a beneficial effect. An *in vitro* study of HL60 (promyelocytic leukaemia) cells treated with decitabine (another hypomethylating agent) identified 160 genes as undergoing demethylation and upregulation (FABIANI *et al.* 2010). The genes identified were often involved in differentiation and development pathways and included *DAPK1* as one of the target genes. Another study using colorectal cancer and leukaemic cell lines showed that the genes demethylated by these agents are both reproducible and non-random (HAGEMANN *et al.* 2011). Currently azacitidine is only licensed for use in haematological malignancies but *in vitro* studies are investigating its possible use in other malignancies.

1.6 Techniques for studying DNA methylation

Techniques for studying DNA methylation have helpfully been described as falling into those useful for "typing" and those suitable for "profiling" methylation (BECK and RAKYAN 2008). The former provide information at single CpG resolution, but are difficult to extend genome-wide and are usually used at specific target loci. They include methylation-sensitive restriction enzyme mapping, bisulfite sequencing, methylation sensitive PCR and pyrosequencing, amongst others. Profiling techniques provide information on a genome-wide scale, such as enrichment of methylated DNA with an antibody to methylcytosine (MeDIP- (WEBER *et al.* 2005) or to the MBD2/MBD3L1 protein complex (MIRA),

followed by hybridisation of the recovered material to a microarray (RAUCH and PFEIFER 2005). Other examples include Restriction Landmark Genomic Scanning (RLGS) (ANDO and HAYASHIZAKI 2006), and the HELP assay (HpaII tiny fragment Enrichment by Ligation-mediated PCR) (KHULAN *et al.* 2006). What these techniques gain in breadth of coverage they lose in resolution. It is possible to combine both resolution and breadth, such as high-throughput sequencing of an entire bisulfite modified genome, as has been done for *A thaliana* (COKUS *et al.* 2008). Importantly, the combination of technical difficulties of mapping back bisulfite-modified sequences to the genome (due to the reduced complexity created by the conversion of Cs to Ts) and the prohibitive cost render this a technique currently unavailable to most laboratories.

While it would be impractical to review all the techniques used to study DNA methylation, I will provide a summary of key techniques below, but must first mention HPLC, and in particular thin layer chromatography. This was one of the first techniques used to study DNA methylation and it is limited to giving a global assessment of DNA methylation levels, as it relies on the hydrolysis of DNA to individual bases and therefore cannot give information on the location of 5mC (OAKELEY 1999). With the emergence of increasingly complex tools able to look at DNA methylation at both individual and global loci, this technique could have expected to be confined to the dustbins of history. Instead, it has made an impressive resurgence with the recent discovery of 5hmC as a key base in development, and HPLC remains the most reliable method of detecting it in tissues, albeit only at a global level.

1.6.1 Methylation typing

1.6.1.1 Methylation-sensitive restriction enzyme analysis

The earliest technique used to assess DNA methylation at individual loci made use of the differential sensitivity of various restriction enzymes to the presence of methylation. For example, HpaII cuts its recognition site (CCGG) only if the CpG is unmethylated. When the fragments generated are resolved on an agarose gel and hybridised with a radioactive probe corresponding to the locus in question on a Southern blot, the size of the bands produced gives a highly

reliable indication of the number and location of methylated CpGs. This technique is extremely sensitive to small changes in DNA methylation and is semi-quantitative, but large amounts of starting material are necessary and it is only informative at the sites of restriction enzyme cutting (OAKELEY 1999). Furthermore, incomplete digestion for a reason other than methylation can lead to false positives. While the use of a specific probe assesses methylation in one particular target, it can be extended to look more genome-wide by using a probe which would hybridise to a common region in repetitive elements.

1.6.1.2 Bisulfite sequencing

The discovery that exposure of DNA to sodium bisulfite at high temperatures and low pH led to a differential reaction in methylated 5mC and unmethylated cytosines has been instrumental to the study of DNA methylation (FROMMER *et al.* 1992). The reaction converts unmethylated cytosines to uracils, but 5mC are protected and remain as cytosines. Following PCR amplification and sequencing, thymines are incorporated in place of uracils. The sequencing reaction will therefore yield CG for methylated CpGs and TG for unmethylated ones in the original sequence. For long referred to as the gold standard, this technique provides methylation information at individual CpGs and the use of nested PCRs in the amplification process improves the sensitivity that would otherwise be lost once the sequence complexity has been reduced to 3 bases. Disadvantages of this technique include fragmentation of DNA due to the harsh treatment conditions, false positive rates from incomplete conversion of cytosines, and bias generated by PCR amplification (30-fold variability between the amplification of fully methylated and fully unmethylated substrates (WARNECKE *et al.* 1997)) and/or subcloning processes. As the technique is only semi-quantitative with the degree of methylation estimated from relative C vs T peak heights, it is possible to improve on the quantitation by subcloning a number (usually 10-20) of molecules and determining CpG methylation from the average of the subcloned sequences. Non-conversion can be sequence-specific (thought to be due to secondary structures) or sporadic (WARNECKE *et al.* 2002).

1.6.1.3 Pyrosequencing

The C vs T differential following bisulfite treatment has been further exploited with the use of pyrosequencing. This technique was initially developed to quantitatively assess SNPs, however bisulfite treatment of DNA effectively creates a C/T SNP. In the pyrosequencing reaction, nucleotides are added sequentially to a DNA template in the presence of polymerase. A cocktail of enzymes uses the phosphate released after incorporation of the dNTP and converts it to light. The intensity of the light emitted is recorded. The height of each of these peaks is used to accurately determine the order of the incorporated nucleotides and hence the sequence of the DNA, reliably estimating relative amounts of C vs T incorporation. This can be successfully used to determine methylation percentages in human DNA from 10ng starting material (DUPONT *et al.* 2004). The advantages of this technique are that it is rapid, highly quantitative, applicable to CpG rich and CpG poor regions and can be multiplexed to large numbers of samples. Unfortunately, it can only sequence short reads at a time, making it time consuming to cover large areas, and not every sequence is suitable for pyrosequencing. It can be quite time-consuming to set up and validate the assay for one target region, but this makes the technique well suited to studying a few regions in many samples rather than many regions in a few samples.

1.6.1.4 MALDI-TOF (Matrix-assisted laser desorption ionization time-of-flight)

Another highly quantitative assay useful for methylation typing is MALDI-TOF mass spectrometry. In this technique, DNA is bisulfite modified and digested with a restriction enzyme, followed by time-of-flight mass spectrometry of the products. The difference in mass created by the C vs T base change in methylated/unmethylated molecules allows the determination of methylation at assessable CpGs. Not all CpGs can be assessed by the technique but it is reported to be highly quantitative and can be carried out on as little as 400pmol DNA (TAMURA *et al.* 1997). Its disadvantages include the limited throughput and high cost (GUPTA *et al.* 2010).

1.6.1.5 Methylation-specific PCR

Another use of the C vs T transition following bisulfite treatment is in the use of methylation specific PCR (FACKELMAYER *et al.*). Here primers are designed to anneal to either methylated or unmethylated fragments and the relative amounts of the products generated give an indication of the amount of methylation at the site of primer annealing (HERMAN *et al.* 1996). This is taken further for more accurate quantification by combining it with RQ-PCR in the MethyLight assay where primers and/or probes cover methylated or unmethylated CpGs (EADS *et al.* 2000).

It must be highlighted that a lot of these techniques are based on the C/T bisulfite reaction and that the discovery that 5hmC behaves like 5mC in that it does not convert to thymine (HUANG *et al.* 2010) raises questions regarding previous studies which would have identified 5hmC as 5mC. There is now a mAb specifically for 5hmC, which may help clarify this issue (GUPTA *et al.* 2010).

1.6.2 Methylation profiling

Some of the techniques described in the typing section can be scaled up for profiling.

1.6.2.1 RLGS (restriction landmark genomic scanning)

One of the first techniques to be scaled up in this way was RLGS (restriction landmark genomic scanning) (HAYASHIZAKI *et al.* 1993). With this technique, a methylation sensitive restriction enzyme (e.g. NotI) is used and then the digests are end-labelled with a radionucleotide and subjected to two-dimensional gel electrophoresis. Only unmethylated (cut) fragments will be labelled. From the intensity of the spots generated after electrophoresis it is possible to quantitatively determine the methylation status of each allele tested. Clearly only alleles with the corresponding RE recognition site will be assessed. Comparison with a bioinformatically generated "virtual RLGS" map covering all possible restriction fragments containing all possible combinations of methylated/unmethylated CpGs allows identification of exact sequences and

determination of which CpGs have gained/lost methylation (SMIRAGLIA *et al.* 2007).

1.6.2.2 RRBS (reduced representation bisulfite sequencing)

Initial attempts to scale up bisulfite sequencing were successful in the application of RRBS (reduced representation bisulfite sequencing)(MEISSNER *et al.* 2005). A restriction enzyme is used to digest the DNA and the products are then size selected (eg 500-600bp), creating a "reduced representation" of the genome, and these fragments are then bisulfite converted and amplified, cloned and sequenced. The entire process creates a 'random' selection of the genome, which is more amenable to large-scale sequencing than whole genome sequencing. It was found to be adequately representative of key areas, as it covered 65% of core promoters, 50% of CGIs and 17% of possible regulatory elements (GU *et al.* 2010). Advantages of this method include small amounts of starting material (30ng) as well as its use on formalin-fixed or paraffin embedded material, making it applicable to clinical samples. A major advantage is its lack of bias for CpG density and when compared to other techniques it was found to be reliable at identifying known DMRs (BOCK *et al.* 2010).

1.6.2.3 MeDIP (Methylated DNA Immunoprecipitation)

With the proliferation of microarrays for various uses came their application to the study of DNA methylation. They were initially combined with affinity-based measures for enriching DNA for methylated regions. The first such technique to be published was MeDIP. This is based on the principle of immunoprecipitating DNA with a monoclonal antibody against methylated cytosines, amplifying the DNA, labelling it and hybridising it to a microarray. In theory, this should provide a global unbiased assessment of the methylation status of all CpGs, albeit not at single CpG resolution. The starting material is sonicated DNA, therefore MeDIP will assess methylation over a range of CpGs present in various fragments. Furthermore this assessment should be quantitative, as the amount of bound:input signal generated should be proportional to the amount of DNA precipitated by the antibody, itself a function of the proportion of methylated CpGs. The initial report suggested that MeDIP was able to fulfill all these criteria

(WEBER *et al.* 2005). It quickly became evident that this technique is biased towards areas of high CpG density and complicated statistical normalisations to correct signal intensity for CpG content had to be devised (DOWN *et al.* 2008). Nevertheless, some comparisons of MeDIP with other techniques have shown it to have low sensitivity and specificity although it was able to pick up some DMRs picked up by other methods (BOCK *et al.* 2010).

1.6.2.4 CHARM (comprehensive high throughput array for relative methylation analysis)

Another technique based on microarray analysis is that of CHARM (IRIZARRY *et al.* 2008b). This is based on fractionation of the genome with McrBC, which cuts only methylated DNA, and hybridisation to a carefully designed array which maximises the number of CpGs covered to those which can give the most reliable information. It must be noted that methylation levels are not assessed at an individual CpG resolution and it is most accurate at assessing methylation in fragments 200-600bp. Combining these variables with careful bioinformatics analysis, the methylation assessed by CHARM reaches 100% sensitivity and 90% specificity (IRIZARRY *et al.* 2008b). Furthermore, the CpGs assayed are not biased towards CGIs.

1.6.2.5 MIRA (methylated island recovery assay)

Another useful technique is MIRA (methylated island recovery assay). This makes use of a combination of MBD2B/MBD3L1 to bind methylated DNA. It is an affinity method without antibodies. The selected methylated fragments are amplified, labelled and hybridised to a microarray and provide methylation information at 100bp resolution (RAUCH and PFEIFER 2010). As its name implies, MIRA focuses on methylated DNA and it has been shown that a loss of methylation of 60-90% must be present before can be detected by MIRA (RAUCH *et al.* 2008).

1.6.2.6 Illumina HumanMethylation 27 BeadChip

Finally, a useful technique to look at a snapshot of DNA methylation genome-wide is the Illumina HumanMethylation 27 BeadChip. This assesses the methylation of 27578 CpGs at key locations in the genome by containing a bead that can bind a methylated CpG or an unmethylated CpG for the same site on bisulfite treated DNA. The fragments are then amplified and the amount of C vs T reaction is detected by the amount of single-base extension that occurs at each of these beads. Detection is based on fluorescence from the labelled ddNTPs.

1.6.3 A new gold standard

With the advent of next generation sequencing, the logical step was to use some of the methods described in the previous section and instead of detecting the enriched DNA by microarray, subject it to high throughput sequencing as has been applied to RRBS, MIRA, and MeDIP for example. Methyl-Cap-seq, which uses MeCP2 to pull down methylated DNA and then produces different eluates of varying CpG densities and subjects them to high throughout sequencing, is another example (BRINKMAN *et al.* 2010). While this approach removes the inherent biases that come with microarrays (e.g. differential hybridisation, dye incorporation biases (KNAPEN *et al.* 2009)), it fails to remove the biases that come with the individual techniques.

A more unbiased use of next generation sequencing is its use for sequencing either part of or the whole genome following bisulfite modification. This was initially carried out in a small *Arabidopsis* genome (COKUS *et al.* 2008) but before long was applied to human cells either on the whole genome (LAURENT *et al.* 2010; LISTER *et al.* 2009) or a targeted part of the genome, but unlike RRBS was directed to particular regions of interest (DENG *et al.* 2009; HODGES *et al.* 2009). The limitations that come from these techniques are mainly related to the cost associated with the required depth of sequencing but also involve significant bioinformatics issues that are becoming clearer as the technique is being used more. These will be reviewed in more detail in the introduction to Chapter 4. Certainly, whole genome bisulfite sequencing seems set to be the new gold

standard for methylation analysis in the same way that Sanger bisulfite sequencing previously was.

1.6.4 The next horizon

And yet is not surprising that even as the latest "new" technology barely comes into routine use, the next level of analysis is being developed. While whole genome bisulfite sequencing may seem like the ideal method for analyzing DNA methylation, it still relies on bisulfite modification and a degree of amplification, which may introduce bias during the library preparation. A true unbiased tool would be one that could detect differences between 5mC and unmethylated C (and perhaps even 5hmC) without need for modification or amplification. This may be the case with nanopore sequencing. This technology relies on the electrical signal that develops as a DNA molecule passes through a pore in a membrane. In this way individual nucleotides, be they canonical or modified by methylation or hydroxymethylation, can be recognised as they pass into the pore. This technique is alleged to be very rapid, cover long reads (>5kb) and not require any PCR amplification. Parameters such as thickness of the membrane and pore diameter affect the resolution and these may be more or less sensitive to modified bases (MIRSAIDOV *et al.* 2010). Nevertheless this technology may prove ultimately useful for unbiased analysis of DNA methylation.

Finally, SMRT (single molecule real time) sequencing is a technique which observes polymerases in real time as they incorporate fluorescently labelled nucleotides complementary to a template. As the nucleotide is added, a pulse of fluorescence determines its identity by colour, duration and interval before the next pulse, and these variables are able to distinguish modified bases such as 5mC and 5hmC. Again the advantages of this technique are that no bisulfite modification or prior amplification is necessary (FLUSBERG *et al.* 2010).

Currently these techniques are still in development and are only able to determine methylation in homogeneous samples, meaning that all of the CpGs at a particularly locus are either fully methylated or fully unmethylated. In most samples studied, however, there will be varied methylation at a particular CpG

due to the variable methylation status of that CpG in the population of cells studied, a problem which will eventually need to be overcome (SUMMERER 2010).

1.7 Chronic lymphocytic leukaemia

1.7.1 The clinical condition

Chronic lymphocytic leukaemia (CLL) is a malignancy characterised by the abnormal accumulation of relatively mature lymphocytes. It is one of the commonest malignancies in the Western World with an incidence of 30/100,000 over the age of 70 years. Although it is not rapidly fatal and patients have a median survival of about a decade, it remains incurable and carries significant morbidities associated with both the disease itself and the treatment to control complications (DESAI and PINILLA-IBARZ 2012).

Clinically, CLL is divided roughly equally into two different prognostic subgroups- patients with a mutated IGVH, who have a median survival of 8 years, and those with an unmutated IGVH who have a median survival of 25 years. These distinct conditions have different gene expression profiles and associations with other prognostic factors, and are reflected biologically by a difference in the behaviour of the B-cell receptor (BCR). In mutated cases, signalling through the BCR invokes anergic and anti-apoptotic responses, while in the unmutated cells leads to downstream activating and proliferating responses (LANASA 2010).

In recent years it has become apparent that CLL often evolves from a pre-malignant condition called MBL (monoclonal B-cell lymphocytosis). As CLL cells have a characteristic immunophenotypic profile, it was noted that a proportion of the normal population (e.g. 0.6%) harbours a small amount of these monoclonal abnormal cells by immunophenotyping (median of 11%), within the context of a normal total lymphocyte count. What remains unclear is what proportion of individuals with MBL will eventually develop CLL (RAWSTRON 2009).

CLL has been notoriously difficult to treat with the emphasis being on symptom control rather than cure. Attitudes to treatment have shifted over recent years with the advent of combination chemotherapy for CLL in the form of FCR: Fludarabine, Cyclophosphamide and Rituximab, an anti-CD20 monoclonal antibody. This combination affords a 3-year disease-free survival of 65% for patients fit enough for treatment (HALLEK *et al.* 2010). Unfortunately, most patients relapse and eventually become refractory to further treatment and it is not clear that there is any overall survival benefit of treatment. Nevertheless, a host of new treatment modalities from cyclin-dependent kinase inhibitors, Bcl-2 inhibitors, and protein kinase C inhibitors to name a few are under investigation in clinical trials (O'BRIEN 2008).

1.7.2 Genetic anomalies in CLL

CLL has a relatively high incidence of familial cases, and yet linkage analysis studies have so far failed to identify causative genetic lesions. The reasons for this have been attributed to low numbers of individuals in affected families, late age of onset and what can only be assumed to be genetic heterogeneity (LANASA 2010). Nevertheless, one family has been identified in which a point mutation leads to *DAPK1* down-regulation (RAVAL *et al.* 2007). A number of acquired mutations have been identified such as in *p53* and *ATM*. Furthermore, recurrent cytogenetic abnormalities are frequently found in CLL. These include 13q14 deletions (found in 55% of cases, site of *ATM*), 11q22-23 deletions (18% of cases) and 17p13 deletions (7% of cases), the latter of which corresponds to a loss of p53 and a very poor prognosis, in particular as >80% of cases are associated with a p53 mutation on the non-deleted allele (DOHNER *et al.* 2000). Conversely, 13q14 deletions are associated with a relative good prognosis.

1.7.3 Epigenetic anomalies in CLL

Interest in the 13q14 deletion was sparked when the lesion in question was found to centre on the deletion of miRNA 15/16 at this locus (CALIN *et al.* 2002). These miRNAs normally down-regulate the anti-apoptotic molecule Bcl-2 and their loss in CLL leads to inappropriate escape from apoptosis. The perturbation of micro RNAs is an example of an epigenetic effect in CLL. Another example would be the finding that *DAPK1* is frequently silenced in CLL due to promoter

methylation (DEBATIN 2007). This provides a direct link between genetic and epigenetic lesions. While there is some debate on the significance of involvement of particular genes by genetic and epigenetic lesions, the finding that some genes are targeted by genetic defects at one allele and epigenetic defects at the other follow Knudson's two hit hypothesis.

CLL was one of the earliest malignancies shown to have DNA methylation abnormalities. Using two different techniques, the overall content of 5mC was shown to be decreased by 26.6% in CLL cells compared to normal (WAHLFORS *et al.* 1992). Hypomethylation of particular regions is critical to CLL. CLL cells are characterised by increased levels of the anti-apoptotic protein Bcl-2. The promoter of this gene was found to be hypomethylated by restriction enzyme analysis in CLL cells. Although there is no mechanistic proof that the change in methylation is responsible for the increased expression, the observation remains interesting in view of CLL cells' dependence on evasion of apoptosis (HANADA *et al.* 1993). The epigenetic effects on Bcl-2 are further compounded because of deranged Aiolos expression. This is a zinc finger transcription factor which is overexpressed in CLL cells and which can induce over-expression of several Bcl-2 family members. In CLL, the *Aiolos* promoter is demethylated and the histones enriched for H3K4me3 compared to normal cells (BILLOT *et al.* 2011). Furthermore, CLL cells have inappropriate activation of *wnt* signalling pathways and the genes which normally counteract this activation have been found to be abnormally methylated in CLL (BENNETT *et al.* 2010).

Like all malignancies, CLL has been shown to be associated with hypermethylation at a number of promoters. RLGS has shown that 4.8% of CGIs are aberrantly hypermethylated in CLL and that these are non-randomly distributed (RUSH *et al.* 2004). As expected some of these targets are specific to CLL (e.g. *GNMT0*) while others are hypermethylated in a variety of malignancies (e.g. *TBX8*). Another study showed 280 new targets of methylation in CLL which were grouped on 4 chromosomes in particular and predicted for IGVH mutational status and Zap70 expression. The expression was lower in methylated genes (TONG *et al.* 2010). Genes which have been shown to be

hypermethylated and silenced are enriched for functions related to apoptosis and cell cycle signalling (PLASS *et al.* 2007).

One study has found that DNMT levels are not affected in CLL while MBD2 and MeCP2 are increased (KN *et al.* 2004).

Some of the DNA methylation anomalies carry significant prognostic weight: silencing of *Zap70* by promoter methylation is associated with an unmutated IGVH and longer median survival (YU 2006). Furthermore, the overall level of 5mC adjusted for age and total white cell count was found to correlate with need for early treatment (YU *et al.* 2007). Equally, the pattern of DNA hypermethylated promoters correlates with CD38 +ve or –ve disease, another prognostic marker (RAHMATPANAH *et al.* 2009) and IGVH mutated and unmutated status (KANDURI *et al.* 2010). *TWIST2* (a bHLH transcription factor) is methylated and silenced in IGVH mutated cases and correlates with good prognosis (RAVAL *et al.* 2005).

Epigenetic lesions are also implicated in progression of CLL. The promoter of the DNA mismatch repair enzyme hMLH1 is frequently hypermethylated in CLL in transformation to Richters, and is associated with microsatellite instability (FULOP *et al.* 2003).

A transgenic mouse model was developed for CLL by introducing *TCL1* (a T-cell specific gene) under the B-cell Ig promoter control. The mice initially develop a non-clonal B-cell expansion followed by a frank B-cell leukaemia. This is associated with promoter hypermethylation at a rate consistent with that reported in human B-CLL (2-5%). The genes whose promoters were hypermethylated were silenced and the methylation changes were found early in the course of the development of the leukaemia. In this mouse model, FOXD3 seems to play a crucial role in the methylation abnormalities, as *FOXD3* was abnormally methylated and the promoters of the hypermethylated genes were highly enriched for FOXD3 binding sites. Knock-down of FOXD3 in an experimental system led to hypermethylation and silencing of its target genes.

While the mouse model is more simplistic than the human disease, being driven by a single genetic lesion, the involvement of DNA methylation abnormalities early in the course of the disease raises interesting questions on the role of DNA abnormalities in primary human B-CLL (CHEN *et al.* 2009).

Given the importance of methylation changes in CLL it might be inferred that the condition might respond to hypomethylating agents. *In vitro* studies have shown that treatment with demethylating agents combined with HDAC inhibitors induce CLL cells to re-express tumour antigens and a functional T cell response which could play a cytolytic role against CLL cells (DUBOVSKY *et al.* 2011). Surprisingly, studies of demethylating agents in CLL remain sparse. One Phase 1 study of relapsed/refractory CLL patients receiving decitabine found that due to toxicity, the limited dose that could be administered was insufficient to lead to changes in DNA methylation (BLUM *et al.* 2010). An HDAC inhibitor alone was trialled in a phase 1 study and did demonstrate genome-wide increase in histone acetylation at tolerable doses (BYRD *et al.* 2005) however this does not seem to have led to further trials.

1.8 Outstanding questions in the field

Despite the remarkable progress that has been made regarding the scope of DNA methylation anomalies in malignancy, several key questions remain outstanding:

(1). Are the methylation changes seen in malignancy causal to the development of the disease or do they result from dysregulated cellular pathways once the uncontrolled proliferation is established?

(2). Do the methylation changes seen in malignancy occur in a random or directed manner, and if the latter, how are these templated?

(3). What is the mechanism by which methylation abnormalities arise in a malignant cell?

Although initial studies suggested that tumour suppressor genes are particular hypermethylation targets in malignancy, it now appears more likely that the genes that are targeted by hypermethylation are already repressed in normal cells. Therefore, it is less likely that methylation is involved in silencing of key genes that would drive cancer initiation and progression. Nevertheless, the studies that have highlighted this key observation are based on the assessment of thousands of loci and reflect general trends in methylation changes occurring at these loci in a population of cells. This does not exclude the possibility that key genes could undergo targeted methylation changes either as part of a general dysregulation of methylation or through a specific process affecting only a few genes, such that these changes could be involved in driving the malignant process. In the same way that "driver" and "passenger" mutations have been used to describe genetic alterations in malignancy, with the former affecting a fraction of the number of the latter, it is possible that the instructive mechanism that seems to target silenced genes for deeper repression by DNA methylation in malignancy represent "passenger" epimutations which result from an aberration in the DNA methylation machinery and that "driver" epimutations do arise from *de novo* methylation and silencing of key tumour suppressor genes. The non-random distribution of methylation changes in malignancy certainly suggest that these occur in a coordinated fashion, but the important observation that there is a high degree of stochasticity in certain regions and that this might reflect a "deliberate" amount of plasticity in epigenetic control to allow more rapid adaptation to evolutionary pressure in tumour development is very intriguing. Yet all of these observations and hypotheses serve only to highlight the lack of answer to the final question that is the mechanism for abnormal methylation in malignancy. No doubt there will be several mechanisms at play, not least of which to explain disparate findings such as hypomethylation in some regions and hypermethylation in others, but most likely even several mechanisms will be responsible for either hyper- or hypomethylation. The tentative links between stem cell epigenetic control and cancer progenitor epigenetic control point in a certain direction, but without providing a specific answer.

1.9 Project plan

The project therefore sought to focus on the key question of the mechanism behind methylation abnormalities in malignancy.

When the project was conceived, genome-wide methylation techniques were still in their infancy and there was little data on unbiased genome-wide methylation changes in malignancy. The project was therefore formulated to start with a descriptive aspect, assessing changes in methylation in one malignancy, chronic lymphocytic leukaemia (CLL). This was to be carried out both globally, using a newly published immunoprecipitation-based technique purported to accurately assess DNA methylation genome-wide albeit at medium resolution, as well as at high resolution in a neutral locus for which there is no selective pressure, namely the alpha globin cluster. The pattern of methylation changes in this region should be representative of changes occurring in other regions of the genome where a functional effect would then provide the cells with a positive or negative advantage. Critically, the lack of selective pressure from the alpha globin locus should allow the assessment of methylation changes in an unbiased way. In parallel, due to the known importance of cross-talk between DNA methylation and histone modifications, an assessment of changes in histone modifications in CLL would also be carried out.

After describing the epigenetic changes in CLL at a global and local level, analysis of these was then expected to shed light into the disrupted mechanisms at play in malignancy and allow the formulation of a hypothesis to explain the widespread changes in DNA methylation.

The data from the descriptive aspect of the project pointed to a mechanism of passive hypomethylation resulting from over-expression of DNA-binding proteins in malignancy leading to their aberrant binding through the genome, resulting in CpGs within their binding sites being sheltered from the DNA methylation machinery. This hypothesis was tested using an EMSA approach followed by DNA affinity chromatography and mass spectrometry in an attempt to identify proteins involved in this process. While this did not lead to the

definitive identification of any such protein(s), certain candidates were identified which could be further investigated and the protocols developed further refined to increase the probability of finding some of these proteins.

Chapter 2. Materials and Methods

2.1 Ethical approval

All samples were obtained under the ethical permission MREC 04.OXB.018 titled "DNA methylation in haematological malignancies" (expiry 30.09.12).

2.2 Chemicals and reagents

Commonly used stock solutions and buffers were prepared as described in Sambrook *et al* (SAMBROOK *et al.* 1989). Chemicals were supplied by Sigma, unless stated otherwise.

2.3 Primers and antibodies

The list of all primers and antibodies is given in Appendix 1.

2.4 Human tissue samples

2.4.1 Normal volunteers

Following informed consent, whole blood was taken by venepuncture from normal volunteers. Alternatively, lymphocyte cones were obtained from the National Blood Service (customer number for ethical approval- T648).

2.4.2 CLL samples

Primary CLL samples were obtained from Dr Anna Schuh and prepared by Maite Cabes using a lymphoprep (axis-Shield) gradient.

2.4.3 Cell isolation

2.4.3.1 Peripheral blood mononuclear cells (PBMC)

PBMC were isolated using a Histopaque 1077 (Sigma) gradient. A 1:1 volume of whole blood and RPMI 1640 complete medium (Gibco) were mixed and layered carefully on top of 1 volume of histopaque. After centrifugation for 30 minutes at 400 RCF, the mononuclear layer at the interface was removed and washed twice in PBS. Cytospins were carried out to verify adequate isolation.

2.4.3.2 Neutrophils

Neutrophils were isolated from the neutrophil/red cell pellet created by the histopaque gradient above. First, red cells were lysed by mixture with lysis buffer (12mM $NaHCO_3$, 155mM NH_4Cl, 0.1mM EDTA) on ice for 15 minutes then centrifuged at 1600rpm 10 minutes, followed by two washes with PBS. Cytospins were carried out to verify adequate isolation.

2.4.3.3 White cell differential isolation by MACS

10^7 cells from the PBMC layer created by the histopaque gradient above were centrifuged at 300 RCF for 10 minutes and the pellet resuspended in 80µl of buffer (PBS pH7.2, 0.5% BSA, 2mM EDTA). 20µl of MACS microbeads (Miltenyi Biotec) of the appropriate specificity were added and incubated on ice for 15 minutes, washed with 1ml buffer and centrifuged at 300 RCF for 10 minutes and resuspended in 500µl of buffer. MS columns were placed on the magnet and washed with 500µl of buffer after which the labelled cell suspension was added. Three 500µl washes were done and the labelled cells retrieved by removing the tube from the magnet and flushing with 1ml of buffer. Selected cells were then washed twice in PBS.

The microbeads used were human CD3 MicroBeads (T cells), CD19 MicroBeads (B-cells) and CD14 MicroBeads (macrophages). The purity of these different fractions were verified by FACS following which CD19 cell purity was then improved by first doing CD14 negative selection followed by CD19 positive selection.

2.4.3.4 CD19 cell selection by DynaBeads

Positive selection of CD19 cells directly from lymphocyte cones was done by adding 25µl of CD19 pan B Dynabeads (Invitrogen) pre-washed with buffer (PBS pH 7.4, 0.1% BSA and 2mM EDTA). These were incubated at 4°C for 30 minutes with gentle rotation and then the beads washed 4x with 1ml of buffer and the magnet, and resuspended in 250µl buffer. The cells were detached from the beads using 10µl Detachabead (Invitrogen) and incubation at room temperature for 45 minutes with occasional gentle pipetting and the beads removed with the

magnet. The cells were then washed in the buffer and centrifuged 6 minutes at 400 RCF followed by two PBS washes.

2.4.4 Assessment of cell purity

2.4.4.1 FACS

Cells were prepared for FACS analysis with assistance from Jackie Sloane-Stanley. FACS was performed in the Molecular Haematology Unit cell sorting facility by Kevin Clark.

2.4.4.2 Cytospins

Approximately 50,000 cells in a volume of 200 μl PBS were spun at 1000 rpm for 5 minutes in a Shandon Cytospin 4. Staining was with Modified Wright's Stain using a Hema-Tek 1000 automated staining machine.

2.5 Tissue culture cell lines

Three CLL cell lines were obtained from DSMZ: JVM2 (ACC-12), JVM3 (ACC-18) and JVM13 (ACC-19). They were maintained in culture by Jackie Sloane-Stanley using 90% RPMI 1640 (Gibco) and 10% fetal calf serum (PAA laboratories) and grown at 37°C with 5% CO_2.

2.6 DNA procedures

2.6.1 Extraction and quantitation of DNA

2.6.1.1 From cell pellets

DNA was isolated by phenol:chloroform extraction. Whole blood was treated with lysis buffer (10mM Tris pH8, 10mM NaCl, 10mM EDTA, 0.5% SDS) and incubated with proteinase K 50μg/ml lysis buffer (Sigma) at 37°C for 4h, adding 1:1000 of RNAseA for the last hour. This was centrifuged at 3400rpm with an equal volume of phenol:chloroform (Fluka phenol saturated with Tris pH8) twice, then an equal volume of chloroform. Precipitation was with 1/10 volume 4M NaCl and double volume 100% ethanol, and the DNA was resuspended with MilliQ water.

Quantitation was carried out on NanoDrop spectrophotometer (Thermo Scientific) or Qubit fluorometer (Life Technologies) as per manufacturer's instructions.

2.6.1.2 From bacterial cultures (MiniPreps)

Plasmid DNA from 2ml bacterial cultures was prepared with the Qiagen MiniPrep kit, an alkaline lysis method using buffers P1 (for resuspension, including RNase), P2 (lysis) and P3 (neutralization) supplied with the kit. DNA was precipitated with isopropanol, and spun down at 13,000 rpm for 15 minutes at 4°C, before washing in 70% ethanol, air-drying and resuspension in 40µl TE at 4°C.

2.6.2 Polymerase Chain Reaction (PCR)

Nested PCR for bisulfite-converted material is described in section 2.10.1.2. For PCR products which were subcloned into DH5α cells, colony PCRs were carried out on the colonies which grew on agar plates. PCR conditions were as per second round of nested PCR but instead of DNA from first round PCR, individual colonies are picked with a sterile tip and added to the PCR tube. The remainder of the conditions were unchanged. The list of primers is in Appendix 1.C.

2.6.3 Quantitative real-time PCR

2.6.3.1 For ChIP assessment

Real time PCR reactions were performed with the Taqman universal PCR mastermix (Applied Biosystems). A typical 25µl reaction contained 6.5µl water, 12.5µl 2×PCR mastermix, 3µl primer/probe mix and 5µl template DNA. The reactions were performed in duplicate with an ABI Prism 7000 sequence detection thermocycler (Applied Biosystems). The list of primers is in Appendix 1.B. Data analysis was performed using the Applied Biosystems software provided.

2.6.3.2 For protein expression analysis

As above but a typical 25µl reaction contained 9.25µl water, 12.5µl 2×PCR mastermix, 1.25µl primer/probe mix (20×) and 3µl template cDNA at 1:10 dilution. cDNA synthesis is described in section 2.7.2. The list of primers is in Appendix 1.A.

2.6.3.3 For MeDIP- Diagenode kit

Primers are supplied in the kit and allow amplification of the positive spiked-in control, the negative spiked-in control, an alpha satellite on chromosome X and GAPDH. 1:4 dilution of the MeDIP reaction is used for RQ-PCR. 25µl reactions are carried out with 12.5µl SYBR Green master mix (ABI), 1µl primers, 6.5µl Sigma water. Amplification cycles are 95ºC 7 min, then 40 cycles of (95ºC 15 sec and 60ºC 1 min), 95ºC 1 min. Results are expressed as $2^{-(CT\ MeDIP-\ CT\ input)}$.

2.6.3.4 For MEDIP- VR protocol

Taqman primers across the 16p locus were designed with Primer Express by other lab members and are available as lab reagents. Taqman primers for the 8 control probes used to quality control the VR protocol were originally designed in the Rakyan lab as SYBR-Green but I re-designed them with Primer Express to be used with the Taqman technology and tested standard curves had R2 >0.997. The list of primers (9066, 4994, 6165, 11851, 6583, 11821, 8804, 5132) is detailed in Appendix 1.B. Reactions were carried out in the same 25µl reactions as above but with the Taqman master mix and 50ng phenol:chloroform extracted DNA per well (as quantified by the NanoDrop). Cycling conditions on the ABI Prism 7000 were 50ºC 2 min, then 95ºC 10 min then 40 cycles of (95ºC 15 sec then 60ºC 1 min). Results are expressed as $2^{-(CT\ MeDIP-\ CT\ input)}$.

2.6.3.5 For targeted capture protocol

The same protocol was used for capture as for testing ChIPs. Two amplicons were used, one inside the target capture area ("103", Appendix 1.B) and one outside ("82", Appendix 1.B). Enrichment was calculated as $2^{\Delta\Delta CT}$ between the inside and outside amplicons.

2.6.4 DNA sequencing

2.6.4.1 Sanger sequencing

Sequencing was carried out by the Core Facility (John Frankland and Tim Rostron) on an ABI 3730 DNA Analyser (Applied Biosystems). The reactions were carried out with 200-500ng of DNA template, BigDye (Applied Biosystems) reaction mix, and 3.2pmol of the relevant sequencing primer. Sequences generated were visualised with Sequencher v4.8. (Gene Codes Corp). Primers used for sequencing were second round PCR primers for bisulfite PCR products (colony PCRs) or the primers supplied in the CloneJET kit for the targeted capture sequencing.

2.6.4.2 High throughout sequencing

DNA was prepared for sequencing by the High Throughput Sequencing Laboratory, Wellcome Trust for Human Genetics using the ChIP-Seq DNA sample prep kit (Illumina) according to the manufacturer's instructions. In brief, 10ng of DNA was end-repaired, and an A base was added to the 3' end of the fragments. Adaptors were ligated so that fragments could be amplified by ligation-mediated PCR. Amplification products 300-500bp were purified by agarose gel electrophoresis prior to sequencing. Sequencing was carried out on Illumina GAIIX at 51 nucleotides in length, paired end.

2.6.5 DNA digestion with restriction enzymes

Restriction digests were carried out with enzymes (Roche or NEB) with the buffers provided at 1x concentration and 1/10 volume of the enzymes and 5-10μg DNA. Incubations were at recommended temperatures, typically 37°C. Digests were visualised by agarose gel electrophoresis, using a 1 % agarose gel with TBE (Tris-borate EDTA) buffer and visualised under UV light after staining the gel in a 1 μg/ml ethidium bromide (Fisher Scientific) solution or 1:50000 GelRed (BioSciences) for 10 minutes. 1μg of 2-log DNA Ladder (NEB) was used as a size marker.

2.6.6 Southern blotting

This was carried out by Helena Ayyub. 5µg DNA was digested with 50U of BglII and HpaII enzymes (Roche) and 1x buffer supplied by the manufacturer and incubated at 37°C for 6h (3h with HpaII alone, followed by the addition of NaCl to 50mM and BglII for a further 3h). Samples were run on a 1% agarose gel, stained in ethidium bromide and visualised under UV light. The gel was soaked in 1M NaOH, rinsed in water, then soaked in neutralization buffer (1M Tris, 3M NaCl, 80ml/L conc. HCl). Blotting was done on a Zeta-Probe GT Genomic membrane (Bio-Rad) in 6x SSC buffer for >4h. The membrane was fixed with UV light and baking at 80ºC for 2h. The probe (plasmid-derived or PCR product depending on which) to the alpha globin region was labelled with α^{32}P (50ng DNA and primers denatured, then labelling buffer with dATP, dTTP, dGTP, followed by ^{32}P -dCTP and the klenow fragment and incubated at 37°C for 10 minutes). The probe was then centrifuged in a Sephadex G50 column. Hybridisation was as follows in high stringency hybridisation buffer (formamide 50%, 20x SSC 3x, Denhardt's 100x, salmon sperm DNA 20µg/ml, Dextran sulphate 6%, SDS 0.5%). The probe was denatured, then added to the membrane and buffer and hybridised overnight. Washes were carried out in 2xSSC at room temperature, followed by 0.1% SDS/0.1x SSC at 65ºC. Visualisation was by photographic film (Kodak) exposure at -70ºC and developed with Kodak processors (Xograph). Quantification was carried out by exposing the membranes to a storage phosphor screen for at least 4h and visualised on a Typhoon phosphoimager (GE Healthcare) using the manufacturer's software for band quantitation.

The probes were all made by Helena Ayyub and are either restriction digests from plasmids (Alpha globin, LINE-1, Alu, TelBam 3.4) or were PCR products (HS-10, pBR11, pXba, Rib7, Rib 11).

2.6.7 DNA ligation

Ligation reactions were carried out with the CloneJET PCR Cloning kit (Fermentas). 5µl DNA from the targeted capture that was cleaved with AcuI, gel extracted and purified with a Qiagen MinElute column was first polished with T4

DNA ligase buffer (1x), 0.4mM dNTP mix, T4 DNA polymerase (3.5U, NEB), Klenow DNA polymerase (5U, NEB), T4 polynucleotide kinase (25U, NEB), incubated at room temperature and cleaned up with Qiagen MinElute kit. This blunt fragment could then be ligated for 30 minutes at room temperature with CloneJET reaction buffer (1x), pJET1.2 vector (50ng) and T4 DNA ligase (5U) and transformed into competent *E. coli* cells.

PCR products for cloning did not need polishing but were directly ligated as above following cleanup with Qiagen PCR purification kit.

2.7 RNA preparation and analysis

2.7.1 Extraction and quantitation of RNA

Approximately 10^7 cells were lysed in 5ml TRI reagent (Sigma), vortexed for 1 minute, left at room temperature for 9 minutes and frozen on dry ice and stored at -70°C. Prior to use, TRI reagent was thawed at room temperature following which 2.5ml of chloroform was added. After vigorous shaking, it was centrifuged at 3200rpm, 4°C for 20 minutes, and the upper aqueous phase transferred to a new tube and 2.5ml of isopropanol was added. After incubation at room temperature for 9 minutes it was again centrifuged at 3200rpm, 4°C for 30 minutes. The RNA pellet was then washed with 5ml of 75% ethanol and centrifuged a further 10 minutes after which the pellet was briefly dried and resuspended in 100µl DEPC-treated water (Ambion). Quantification of RNA was on NanoDrop spectrophotometer (Thermo scientific) and RNA was aliquoted and stored at -70°C.

2.7.2 cDNA synthesis and reverse transcription PCR

RNA was transcribed to cDNA using the high capacity cDNA reverse transcription kit (ABI) according to the manufacturer's instructions. In brief, an enzyme mix was prepared with 2µl 10×RT buffer, 0.8µl 25×dNTP mix, 2µl 10×RT random primers, 1µl multiscribe reverse transcriptase, and 4.2µl nuclease-free water as well as 1µg total RNA in 10µl nuclease-free water. This was incubated in a thermocycler at the following temperature profile: 25°C for 10 minutes, 37°C

for 120 minutes, 85°C for 5 minutes. An RT-ve reaction was always checked alongside the cDNAs when RT-qPCR was carried out.

2.8 Protein preparation and analysis

2.8.1 Nuclear extracts

Nuclear extracts from cell lines grown in culture were prepared according to Dignam *et al.* (DIGNAM *et al.* 1983) with a modification from Osborn *et al.* (OSBORN *et al.* 1989). Briefly, 10^7 cells are washed in cold PBS and resuspended in buffer A, containing 10mM HEPES-KOH, 1.5mM $MgCl_2$ and 10mM KCl with 0.5mM DTT and 1x complete protease inhibitor cocktail (Roche, Switzerland), then incubated on ice for 10 minutes prior to Dounce homogenization. Efficiency of lysis was verified by Trypan blue exclusion. If lysis was sufficient, the samples were transferred to buffer C containing 20mM Hepes-KOH, 25% glycerol, 420mM NaCl, 1.5mM $MgCl_2$, 0.5mM DTT 1x complete protease inhibitor cocktail. After 30 minutes on ice, they were centrifuged 2 min at 13,000 rpm and the supernatant used as a nuclear extract or aliquoted and frozen at −80°C. However, using this protocol lysis was often very poor (30-50%) therefore the protocol was changed to include 0.1% NP40, at which point lysis was consistently 100%.

2.8.2 Quantitation of protein

Protein concentrations were estimated with the BioRad Microplate Protein Assay (BioRad 500-0114) or the Qubit fluorometer (Life Technologies) according to the manufacturer's protocol. Serial dilutions of BSA were used as standards.

2.8.3 Coomassie and silver stain

Nuclear protein extracts were denatured prior to electrophoresis in NuPAGE reducing agent at 70°C for 10 minutes and loaded into the wells of precast NuPAGE 4-12% bis-tris polyacrylamide gels (Invitrogen) and run in MOPS (3-(N-morpholino)propanesulfonic acid) buffer at a constant voltage of 200V for 50 minutes. A prestained marker ladder, PageRuler (Fermentas) was loaded for size determination.

Coomassie blue stain was carried out with PageBlue (Fermentas). After three warm washes in water (placing the gel in 100ml water and microwaving on high power for 1 minute), 20ml of PageBlue Protein Staining Solution was used to cover the gel and microwaved 30 seconds on high power, after which the gel was gently agitated for 20 minutes and washed in water for 5 minutes.

Silver staining was kindly carried out by Monika Kowalczyk. The gel was incubated in a fixing mixture made up of 40% ethanol, 10% acetic acid and 50% water for 1 hr and then washed in water for at least 30 min. The gel was then sensitised with 0.02% sodium thiosulfate for exactly 1 min then washed three times for 20 seconds each in water. The gel was then incubated for 20 min in 4ºC cold 0.1% silver nitrate solution and then washed again three times for 20 seconds each in water. In a new tray, the gel was washed once more in water for 1 minute and developed in 3% Na_2CO_2 0.05% formaldehyde. Staining was terminated by addition of 5% acetic acid for 5 min.

2.8.4 Electro Mobility Shift Assays (EMSAs)

The protocol was adapted from Crossley *et al.* (CROSSLEY *et al.* 1996). EMSAs were carried out by Helena Ayyub and me. EMSAs were performed using 3-5µg of nuclear protein extract. Probes were annealed by mixing 20µM of each forward and reverse oligonucleotides, 0.1M Tris 0.5M NaCl and 1mM EDTA. After incubation at 95°C for 5 minutes, the probes were allowed to cool to room temperature in the heating block removed from heat source. Standard ethanol precipitation followed and the probes resuspended in Qiagen elution buffer. 500ng of double-stranded oligonucleotide probes were then labelled with 2U Klenow fragment of DNA polymerase I and [α-^{32}P]dCTP (Perkin Elmer) and 1x buffer H (Roche) at 37°C for 15 minutes. Unlabelled nucelotides were removed by spinning through a G25 column. In total, 1ng of labelled probe was used in each reaction. Binding was carried out for 20 min at 4°C in 30µl with 10mM HEPES (pH 7.8), 50mM potassium glutamate, 5mM $MgCl_2$, 1mM EDTA, 1mM DTT, 5% glycerol, 1mg poly dI-dC, and 10mg BSA.

For competition experiments, reaction mixtures were preincubated for 15 min with an excess of unlabelled oligonucleotides (2, 4 and 10-fold) prior to the addition of radiolabelled probes. Oligonucleotide sequences are shown in Appendix 1.D. Samples were electrophoresed through a 6% non-denaturing polyacrylamide gel (acrylamide:bisacrylamide ratio, 19:1) at 220 Volts in 0.5× TBE for 2.5 h at 4°C. After electrophoresis, the gels were dried and subjected to autoradiography.

2.8.5 DNA affinity chromatography

The protocol was adapted from Mittler *et al.* (MITTLER *et al.* 2009b).

2.8.5.1 Preparation of double-stranded oligos

Single-stranded oligonucleotides (biotin-labelled in the sense strand) with a spacer sequence and a PstI restriction site as shown in Appendix 1.H are reconstituted at a concentration of 50nmol/ml in annealing buffer (20mM Tris-HCl, pH 8.0, 10mM $MgCl_2$, 0.1M KCl) and annealed at a 3:4 ratio of the biotinylated sense strand to the antisense strand, heated for 5 min at 90°C and gradually ramped down to 65°C over 10min, incubated at 65°C for 5min, and the thermoblock switched-off and allowed to cool to room temperature.

2.8.5.2 Preparation of columns

~250 pmol biotinylated dsDNA was diluted in buffer DW (20mM Tris-HCl, pH 8.0, 2M NaCl, 0.5mM EDTA, 0.03% NP-40) and incubated at room temperature on a rotary wheel for 3 hrs with 1mg of Dynal MyOne (Invitrogen) streptavidin magnetic beads previously equilibrated with two washes TE buffer containing 0.01% (v/v) NP40 and two washes of buffer DW. After the binding step the beads were washed once in TE containing 0.02 % NP40, three times in DW buffer and resuspended in 0.1 ml of buffer DW.

2.8.5.3 Blocking of the beads

Prior to use, the columns were incubated on a rotary wheel for one hour at room temperature in blocking buffer (20mM Hepes, pH7.9, 0.05mg/ml BSA, 0.05 mg/ml glycogen, 0.3M KCl, 0.02% NP40, 2.5mM DTT, 5mg/ml

polyvinylpyrrolidone) and then washed in 1x restriction endonuclease buffer (New England Biolabs buffer 3) containing 0.02% NP40 and twice with buffer G (20mM Tris-HCl, pH7.3, 10% (v/v) glycerol, 0.1M KCl, 0.2mM EDTA, 10mM potassium glutamate, 0.04 % NP40, 2 mM DTT, 1x complete protease inhibitors).

2.8.5.4 Preparation of the extracts

Thawed nuclear extracts were adjusted to 10mM potassium glutamate and diluted with one volume of buffer G to which is added 0.2mg/ml of poly dIdC and the extracts pre-cleared with equilibrated MyOne streptavidin magnetic beads by rotating one hour at 4°C, after which the beads were removed with a magnet.

2.8.5.5 Incubation and washes

Pre-cleared extracts were then incubated for 3h on a rotating wheel at 4°C with the blocked beads (columns) containing the biotinylated oligos at a final concentration of 0.7mg/ml. The beads were then washed four times with buffer G (1.8ml/mg beads) and once (0.9ml/mg beads) with restriction endonuclease buffer (NEB buffer 3; adjusted to 0.02% (v/v) NP40, 2.5% (v/v) glycerol, 1 mM DTT, 1x complete protease inhibitors) and resuspended in this buffer at a concentration of 3.75mg/ml.

2.8.5.6 Elution

A PstI digest was carried out by incubation at 25°C in an Eppendorf thermomixer (at 1400 rpm). This step was repeated once and the beads removed magnetically.

2.8.6 MALDI-TOF

This was carried out by the Core Facility at the WIMM (Weatherall Institute of Molecular Medicine). Bands of the appropriate size on the silver stain were cut under sterile conditions. Following in-gel tryptic digestion and Zip Tip purification and elution, the samples were spotted on a MALDI plate on an alpha-cyano matrix. The samples were then subjected to MALDI-TOF/TOF (UltraflexTM, Bruker Daltonics). Data was analysed with a Mascot interface software.

2.9 Chromatin preparation and analysis

2.9.1 Chromatin immunoprecipitation

Chromatin immunoprecipitation was performed with the Millipore ChIP Assay Kit (Millipore) according to the manufacturer's instructions. Crosslinking of 10^7 cells was done with 0.4% formaldehyde in PBS for 10 mins at 37°C and then quenched with 1M glycine. Following this were two washes with PBS containing 1x complete protease inhibitor cocktail (Roche). Cells were then lysed in SDS lysis buffer and sonicated to an average size of 500-1000bp with a Diagenode Bioruptor and typically frozen at -70°C. Prior to use, 100µl of sonicated chromatin was thawed and diluted 1:10 in ChIP dilution buffer and 10µl of the diluted material taken as input. The chromatin was pre-cleared with protein A : agarose beads, after which immunoprecipitation was carried out at 4°C overnight with the appropriate antibodies (Appendix 1.F). The following day protein A : agarose beads were added for one hour following which washes were performed with high salt buffer, low salt buffer, LiCl buffer and TE buffer. This was followed by elution of bound chromatin fragments, reversal of crosslinks with 0.4 M NaCl at 65°C, proteinase K digestion and purification of DNA by Phenol:Chloroform extraction. Enrichment of immunoprecipitated DNA was quantified by RT-qPCR as a percentage of input DNA with Taqman probes.

2.9.2 Labelling and hybridisation to microarrays

Immunoprecipitated DNA and corresponding input DNA were separately mixed with 60µl random primer solution (BioPrime labelling kit; Invitrogen) and made up to 130µl with water. DNA was denatured in a heat block at 100°c for 10 min. The following reagents were then added on ice: 15µl 10×dNTP mix (Amersham; made up as C 1µl, A 2µl, T 2µl, G 2µl), water 93µl), 1.5µl Cy3 (ChIP DNA) or Cy5 (input DNA) labelled dCTP (1mM; Amersham), 3µl Klenow fragment (Bioprime labelling kit). The resulting mixture was incubated overnight in darkness and stopped by the addition of 15µl stop buffer (Bioprime labelling kit). For each labelling reaction, labelled DNA was cleaned up using Illustra G-50 microspin columns (GE healthcare) according to the manufacturer's instructions.

Following labelling, a hybridisation mix was prepared as follows: 160µl Cy3 labelled DNA, 160µl Cy5 labelled DNA, 135µl human cot-1 DNA (Invitrogen) 55µl 3M NaoAc pH 5.2, 6µl yeast tRNA. The mix was precipitated by the addition of 1ml 100% ethanol and the pellet washed with 70% ethanol before drying. The dried pellet was resuspended in 150µl hybridisation buffer (5ml 20×SSC, 25ml formamide, 5ml 100mM Tris-Hcl pH 7.4, 2.5g Dextran sulphate, 50µl Tween 20, 15ml HPLC water) and allowed to dissolve at 65°c for 1 hour. DNA was subsequently denatured for 10 min at 100°C in darkness.

Samples were loaded onto a custom microarray with probes covering the terminal 500kb of human chromosome 16 (Sanger center). Hybridisation and washing was performed using a Tecan Hybridisation station according to the manufacturer's instructions and including the following wash solutions were used: Channel 1: 2×SSC, 0.1% Tween 20; Channel 2: PBS, 0.05% Tween 20, Channel 4: 0.1% SSC; Channel 6: HPLC water.

Scanning was performed on a ScanArray Gx Plus (Perkin Elmer). The .tiff files generated were analysed with the NimbleScan software to yield log2 ratios of bound:input. A custom-designed perl script (courtesy of Dr Jim Hughes) translated the .gff files generated into .wig files which can be uploaded onto the UCSC genome browser as well as our in-house 16p browser.

2.10 Methylated DNA analysis

2.10.1 Bisulfite sequencing

2.10.1.1 Bisulfite conversion

Bisulfite was modified with the EZ DNA Methylation-Gold kit (Cambridge Biosciences). 250ng of DNA was incubated with the CT conversion reagent at 98ºC for 10min then 64ºC for 2.5h. After desulphonation, samples were washed with the columns as per manufacturer's instruction and retrieved in 10µl elution buffer.

2.10.1.2 Nested PCR

Bisulfite modified DNA was amplified by nested PCR as follows. Bisulfite-specific primers were designed using the Primer 3 0.4.0v software: http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi. These were kindly designed by Supachai Ekwattanakit so that one set of external primers was used for the first round, followed by one set of internal primers for the second round, some of which partially overlap the external primer sequences. Each amplicon was tested by me on bisulfite treated and untreated DNA to ensure only bisulfite treated DNA amplified and that the product of each reaction was a single good quality band and that the sequences genereated were those expected. The PCRs were done with the HotStarTaq from Qiagen in 25µl total reaction for the first round: 1x PCR buffer, 200mM each dNTP, 1mM each primer, 2.5U polymerase and 2µl template from bisulfite modified reaction. Second round PCR was carried out in 50µl with 5µl of first round product with all other reagents at the same concentration as for the first round. PCR conditions were 95ºC 15 min, then 30 cycles of 95ºC 1 min, annealing temperature 2 min, 72ºC 3 min, followed by 72ºC 10 min final extension. 5µl of the second round product was electrophoresed in a 1% agarose gel and visualised with ethidium bromide to ensure there was only one specific band of the correct size. Products of the second round of nested PCR were cleaned up with Qiaquick PCR purification kit (Qiagen) following manufacturer's instructions.

2.10.2 MeDIP (methylated DNA immunoprecipitation)

2.10.2.1 Weber protocol

Genomic (RNAse treated) DNA was sonicated with a Diagenode Bioruptor to create fragments <500bp as seen on 1% agarose gel electrophoresis. 5µg of sonicated DNA was used for the MeDIP reaction, saving 10ng as input. The IP (immunoprecipitated) fraction was denatured for 10min at 95°C and immunoprecipitated for 2h in a 500µl total reaction of IP buffer (10mM sodium phosphate, 140mM NaCL, 0.05% Triton X-100) at 4°C with 10µl of 5-methylcytosine antibody (mouse monoclonal, Diagenode). Washed M-280 magnetic sheep anti-mouse-IgG beads (Dynal Biotech/Invitrogen) resuspended

in 30µl of IP buffer were then added to the incubation for a further 2h. After washing with 700µl of IP buffer, the samples were treated with 70µg proteinase K for 3h at 60ºC. Precipitated DNA was then isolated by phenol:chloroform extraction, washed in 70% ethanol and resuspended in nuclease-free water.

2.10.2.2 VR protocol

This is an adaptation of the Weber protocol with a few amendments by Dr Vardhman Rakyan. 1.5µg of sonicated material was used, and the concentrations of both the antibody (0.5x) and the beads reduced (0.2x). The beads were washed in IP buffer rather than PBS/BSA. Following the MeDIP and the washes, incubation with proteinase K was at 55ºC for 2h. Instead of phenol:chloroform extraction, the recovered DNA was cleaned up using the DNA Clean and Concentrator kit (Zymo Research). Following whole genome amplification (WGA, see below), RQ-PCR was carried out with 8 control probes to verify the enrichment.

2.10.2.3 Diagenode protocol

The MagMeDIP kit from Diagenode (Liege, Belgium) was used. Although the composition of the buffers is proprietary, the protocol is as follows, briefly. DNA was isolated using the GenDNA isolation buffer to which proteinase K is added and incubated at 50ºC overnight. Phenol:chloroform:IAA (Fluka) extraction was performed and the precipitation carried out in GenDNA precipitant and 100% ethanol. Samples were washed in 70% ethanol, dried and resuspended in TE buffer. RNAse treatment was then carried out and the samples re-extracted. The DNA was sonicated and visualised on a 1% agarose gel. Immunoprecipitation was carried out in IP buffer with 1µg of sonicated DNA and 1.5µl of positive and negative control DNA spiked in. Following denaturation, 10% of this was set aside for input. Magnetic beads and antibody were added and incubated at 4ºC overnight. Following this, 6 washes were performed and DNA recovered in DNA Isolation Buffer (DIB) which was incubated at 55ºC for 15 min then 100ºC for 15 min. The input samples were also taken through the DIB step. The supernatant was then removed and used for RQ-PCR. However, I found this was then

unsuitable for use in WGA and required standard phenol:chloroform extraction for WGA to function.

2.10.2.4 Whole Genome Amplification (WGA)

This was carried out using the GenomePlex kit from Sigma. Manufacturer's instructions were followed.

2.10.2.5 Labelling/hybridising/scanning

2.10.2.5.1 In-house

1µg of WGA amplified DNA was mixed with random 9mer primers (Cy-3 labelled for the input and Cy5-labelled for the MeDIP) and VWR water. After denaturation, 50x dNTP mix and klenow fragment were added and incubated at 37°C for 2h. EDTA, 5M NaCL and isopropanol were added and after centrifugation the pellet washed in 80% ethanol. After resuspension in VWR water, Cy3 and Cy5 incorporation were measured by NanoDrop Spectrophotometer. 13µg of MeDIP and input samples were then combined and hybridisation buffer and Cy3 and Cy5 50mer oligos added as controls. After denaturation, the samples were loaded into the MAUI hybridisation machine (BioMicro Systems) with Nimblegen ENCODE arrays and hybridised for 16-20 hours. Following this, arrays were washed with Nimblegen buffers I, II, III and dried.

Scanning was performed on a ScanArray Gx Plus (Perkin Elmer) and the PMTs (laser intensity) chosen so as not to yield too many oversaturated probes. The .tiff files generated were analysed with the NimbleScan software to yield log2 ratios of bound:input. A custom-designed perl script (courtesy of Dr Jim Hughes) translated the .gff files generated into .wig files which could be uploaded onto the UCSC genome browser as well as our in-house 16p browser.

2.10.2.5.2 Nimblegen hybridisation

For the 10 MeDIPs carried out with the Diagenode protocol, 4µg of WGA amplified material was sent to Nimblegen for labelling and hybridising to ENCODE (HG18) arrays. The raw data generated was sent to Dr Thomas Down at the Gurdon Institute for analysis with the Batman algorithm. This translates signal intensity from the array into percentage methylation. Thomas Down also

interrogated the data for differentially methylated regions between samples and used a custom java script to assign these to annotated genes.

All the remaining data analysis was done on Microsoft Excel with the exception of the histograms of the signals generated by scanning a slide at 3 different laser intensities (Figure 3.18B), which was generated in MatLAB 7.9.

2.10.3. Targeted capture

2.10.3.1 Sonication

Sonication of 5µg DNA in a total volume of 100µl EB with the Covaris Adaptive Focus Acoustics (K bioscience) was done for 90 seconds, duty cycle 5%, intensity 3, cycles/burst 200 and frequency sweeping, which are the conditions for shearing at 500bp. Sonication was confirmed by visualisation on a 1.5% agarose gel. The sample was then dried in a SpeedVac lyophiliser (Christ) in an Eppendorf LoBind tube to a final volume of 15µl (60°C for 30 min). Quantitation was with NanoDrop.

2.10.3.2 Capture

5µg sonicated DNA and 1ng biotinylated oligo library in a total volume of 15µl water were denatured 95°C for 5min and slowly annealed at 65°C 15min, following which 15µl of 2x hybridisation buffer (1.5 M NaCl, 40 mM Sodium Phosphate buffer (pH7.2), 10 mM EDTA (pH 8.0), 10x Denhardt's, 0.2% SDS) was added and incubation carried out at 65°C shaking on a ThermoMixer Comfort (Eppendorf) overnight.

50µl streptavidin coated Dynabeads M-280 (Invitrogen) per reaction were washed twice with 500µl bead binding buffer: (10 mM Tris-HCl (pH7.5), 1 mM EDTA, 1 M NaCl) and resuspended in 150µl buffer. This was added to the hybridisation reaction tube and incubated 30min at room temperature. One wash at room temperature in 1ml wash buffer (1x SSC, 0.1% SDS) was followed by another three at 65ºC . Beads were resuspended in 50µl 0.1M NaOH (freshly made), incubated at room temperature for 10min then the supernatant transferred to 50µl 1M tris-HCl (pH 7.5). The eluted, selected material was

desalted using Illustra ProbeQuant G-50 columns. Samples were combined to total of ~100µl.

Enrichment levels were calculated by RQ-PCR (see section 2.6.3.5).

2.10.3.3 Bisulfite conversion

MethylXceed Kit (Human genetic signatures) was used, following manufacturer's instructions, apart from a modification of the cycling conditions to include spikes to 99°C to ensure the fragmented DNA remains denatured: 80°C (10min), 99°C (30sec), 80°C (10min), 99°C (30sec), 80°C (10min), 99°C (30sec), 80°C (15min). Furthermore, samples were pooled prior to addition of 0.5µl of cRNA (1µg/µl, Qiagen) at step 7 instead of Salmon sperm DNA. Elution was carried out in 2x20µl of pre-warmed reagent 5 and incubated at 95°C for 20 minutes.

2.10.3.4 Random Primer Amplification (RPA)

The protocol was adapted from Lieb *et al.* (LIEB 2003).

Round A: Two rounds of DNA synthesis were carried out using a partially degenerate primer (RG3A see Appendix 1.E) at 4µM primer, 4U Sequenase (Sequenase version 2.0, 13U/µl, GE healthcare) and Sequenase reaction buffer as supplied by the manufacturer. After 10min at 94°C, the reactions were placed directly on ice and further Sequenase buffer with 3mM dNTPs (GE healthcare), 0.1M DTT (NEB) and 0.75µg BSA (NEB) and 4U Sequenase were added to each tube and incubated at 37°C 10min, 94°C 2min, 8°C 2min (at which point a further 4U Sequenase was added), 12°C 1min (ramp to 37°C in 8 min). The PCR product was diluted to 50µl by adding 35µl of 1x TE (Fluka) to each PCR tube.

Round B: 5µl of the diluted template from Round A was used to which is added 10µl AmpliTaq buffer (no $Mg^{2}+$), 1µl dNTPs (25mM GE healthcare), 1µl RG4 primer (Appendix 1.E) (50µM), 0.5µl Amplitaq Gold (5U/µl Roche), 0.2µl PWO (5U/µl Roche), 8µl $MgCl_2$ (25mM Roche), H_20 to a total of 100µl. PCR cycles were 94°C 5 min, 30 cycles (92°C 30sec, 40°C 30sec, 50°C 30sec, 72°C 1min).

After checking 10µl of the product on a 1.5% agarose gel, the remainder was purified through a PCR purification column or Minelute PCR purification depending on yield.

2.10.3.5 AcuI digestion

To remove the adapter sequence, purified DNA was incubated with AcuI (NEB) 50U, Buffer 4 1x, SAM 40µM and digested approximately 90min at 37°C. 10µl was run alongside uncut control on a 1.5% agarose gel.

2.10.3.6 Gel purification

Once digestion was complete, the entire products from the digest were run on 1% low melting point agarose in TAE. Fragments above 300bp were extracted and then cleaned up with Qiagen gel purification Minelute column, eluting in 10µl of Qiagen elution buffer.

2.10.4 HPLC analysis for 5hmC levels

This was carried out in the laboratory of Dr Skirmantas Kriaucionis (The Ludwig Institute for Cancer Research, Oxford). The protocol was adapted from Quinlivan *et al.* (QUINLIVAN and GREGORY 2008). In brief, 100µg DNA was hydrolysed overnight in buffer A containing 100mM NaCl, 20mM $MgCl_2$, 20mM Tris pH7.9, to which was added the digest mix of 1000U/ml Benzonase, 600mU/ml phosphodiesterase I, 80U/ml Alkaline phosphatase, 36µg/ml EHNA (erythro-9-(2-Hydroxy-3-nonyl)adenine) hydrochloride and 2.7mM deferoxamine. HPLC was carried out using ZORBAX Eclipse Plus Rapid resolution HD C18, 2,1cmx1,5mm (Agilent). Amicon Microcon centrifugal filter units (10kDa cut-off, Millipore) were used to remove enzymes. The hydrolysate was SpeedVac lyophilised and resuspended in 25µl of 100mM acetonitrile pH6 and injected into the HPLC system (Agilent 1290 Infinity). The run conditions were: 100mM ammonium acetate and a progressive gradient between 1.8-100% of 40% acetonitrile with a flow rate of 0.4ml/min flow rate. The signal was detected at 260nm, 280nm, 254nm and the data analyzed at 260nm.

2.11 Bacterial procedures

Transformation of competent cells

Transformations were done using Library Efficiency DH5α Competent cells (Invitrogen 18263-012). 50μl competent cells with 5μl of CloneJET ligation reaction (Fermentas) were mixed on ice for 30min in a 14ml Falcon polypropylene tube, followed by heat shock at 42°C for 45sec. 500μl SOC medium (supplied with Invitrogen 18263-012) was added and shaken at 37°C for 45 minutes before plating on agar plates with the appropriate antibiotic (40μg/ml ampicillin). Bacterial colonies were grown on agar plates at 37°C overnight. The next day, individual colonies were picked for colony PCR or the bacteria expanded in started cultures. A single streaked bacterial colony was expanded in 4ml LB medium (10g Tryptone, 5g yeast extract, 10g NaCl/L) supplemented with 40μg/ml ampicillin and incubated overnight at 37°C with vigourous shaking (200rpm). The following day, DNA was extracted using the MiniPrep kit (see section 2.6.1.2).

2.11 Bioinformatics

Bioinformatic analysis of the targeted high throughput bisulfite sequencing data is described in detail in Chapter 4, as optimization of the bioinformatics pipeline was an integral part of the work carried out. Processing of the data received from sequencing until input into Seqmonk was done by Dr Nicki Gray (MRC Molecular Haematology Unit, Weatherall Institute for Molecular Medicine). The Bismark software was developed by Dr Felix Krueger (Babraham Institute, University of Cambridge). Discussion and decisions regarding how to alter the analysis was done between me, Richard Gibbons, Nicki Gray and Ricarda Gaentzsch.

Bioinformatic analysis of the MeDIP data was carried out by Dr Thomas Down (Gurdon Institute, University of Cambridge) as described in Chapter 3.

High throughput sequencing reads from ChIP-Seq were mapped by Nicki Gray using Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) to the human

genome (HG18 build). Reads with more than 3 mismatches to the target genome were excluded from subsequent analysis. Peak calling was carried out using MACS1.4 and the peaks were filtered and viewed using Multi-Image Genome (MIG) -a software tool written by Dr Simon McGowan and Dr Jim Hughes (Computational Biology Research Group, WIMM).

Analysis of the distribution of CpGs was carried out using UCSC's Galaxy tool: http://main.g2.bx.psu.edu/ .

Motif analysis was carried out with the MEME suite software tool: http://meme.sdsc.edu/meme/cgi-bin/meme.cgi and transcription factor binding analysis was carried out with TOMTOM: http://meme.nbcr.net/meme/cgi-bin/tomtom.cgi and TESS: http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=SEA-FR-Query.

2.13 Statistics

All error bars on the graphs represent standard deviations.

Extreme changes of methylation were calculated by determining all values beyond two standard deviations at either extreme.

Observed to expected (O:E) ratios were calculated by dividing the prevalence of an observation in the sample population by the prevalence of the same observation in the entire population (e.g. proportion of hypomethylated CpGs found in exons/proportion of all CpGs found in exons). Statistical significance of these ratios was calculated using the binomial probabilities test.

Pearson's correlation analysis was performed to determine the association between different variables.

Student's t-test was used to compare the significance in the difference in mean methylation levels between two groups.

Chapter 3. DNA methylation abnormalities in CLL

3.1 Introduction

The initial aim of the project was to describe methylation anomalies in CLL at both a local (alpha globin locus) and global (genome-wide) levels. The choice of CLL as the malignancy on which to focus these studies stemmed from previously described changes in methylation associated with this malignancy (WAHLFORS *et al.* 1992), and the availability of homogeneous cells from patients' peripheral blood. The studies would be carried out using Southern blot analysis after methylation-sensitive enzyme cleavage, traditional bisulfite sequencing and an immunoprecipitation based method for genome-wide methylation analysis, MeDIP (Methylated DNA Immunoprecipitation).

3.2 Preliminary findings

3.2.1 Methylation abnormalities at 16p in malignancies

A rapid assessment of methylation can be obtained by coupling methylation-sensitive restriction enzymes (RE) with Southern blotting. This is based on the finding that some REs, such as HpaII, cut at their recognition site (CC/GG) only if the CpG is unmethylated. When the restriction fragments generated are resolved on an agarose gel, transferred to a membrane and hybridised with a radioactive probe corresponding to the locus in question, the size of the bands produced gives a highly reliable indication of whether or not the CpG dinucleotide is methylated. Using this technique, DNA from the region around alpha globin (16p13.3) was studied in the peripheral blood of patients with a variety of haematological malignancies and that of normal subjects. Figure 3.1A illustrates the pattern of methylation in nine normal individuals and nine patients with CLL (chronic lymphocytic leukaemia). While normal individuals have a uniform pattern of methylation, CLL patients show hypomethylation, as evidenced by the presence of smaller bands from more frequent cutting. The patterns in CLL differ not only from normals but also between patients, suggesting that the CpGs that become hypomethylated in CLL are not identical between patients.

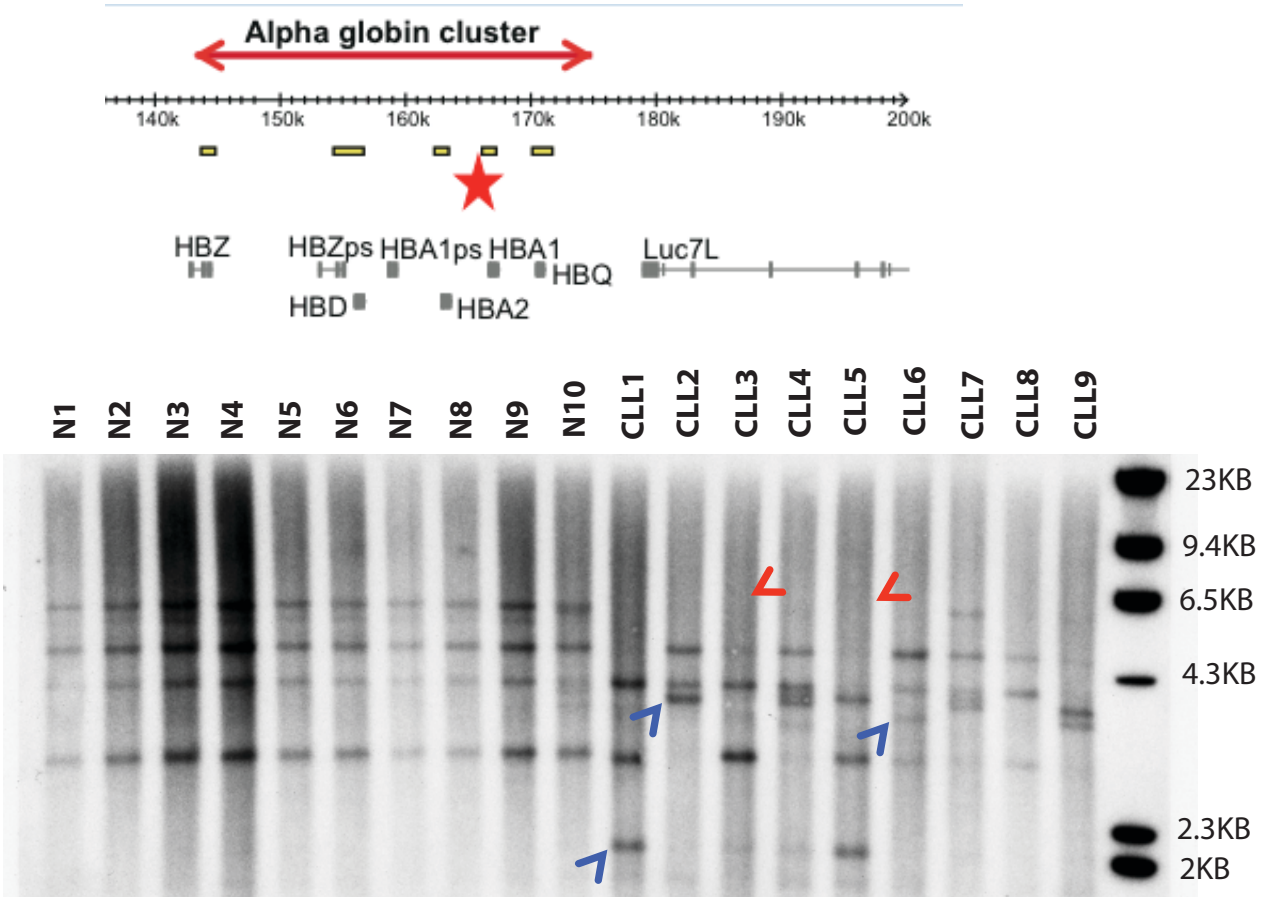


Figure 3.1A. Southern blot radiograph of nine normal individuals and nine CLL patients. The DNA was digested with BglII and HpaII, electrophoresed in an agarose gel, blotted to a membrane and hybridised with a probe corresponding to a PstI fragment at the alpha globin CpG island. N= normal individual, CLL= chronic lympocytic leukaemia patient. Red arrowheads show loss of large bands and blue arrowheads gain of smaller bands, indicating hypomethylation. Red star illustrates location of the probe. Marker is Lambda/HindIII (NEB).

Figure 3.1B. Prevalence of methylation abnormalities observed at the alpha globin locus in normal individuals and patients with a variety of haematological malignancies. Anaemia= patients under investigation for anaemia, MDS= myelodysplastic syndrome, ATMDS= Acquired thalassaemia in MDS, CML= chronic myeloid leukaemia, ALL= acute lymphoblastic leukaemia, CLL= chronic lymphocytic leukaemia, NHL= Non-Hodgkin's lymphoma.

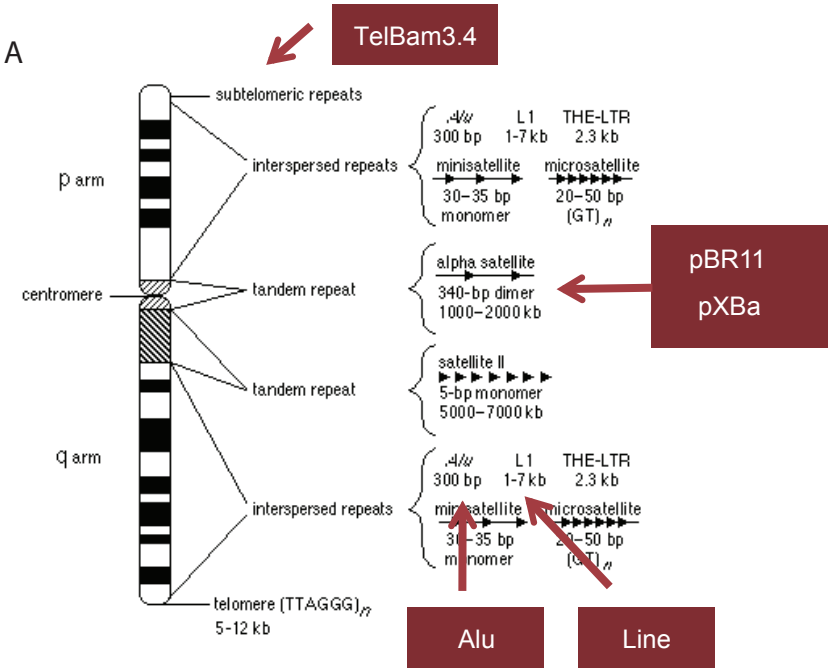
Diagnosis	% Abnormal Methylation
Normal	6 (n=49)
Anaemia	7 (n=42)
MDS	51 (n=56)
ATMDS	77 (n=13)
CML	35 (n=19)
ALL	100 (n=1)
CLL	100 (n=14)
NHL	84 (n=39)

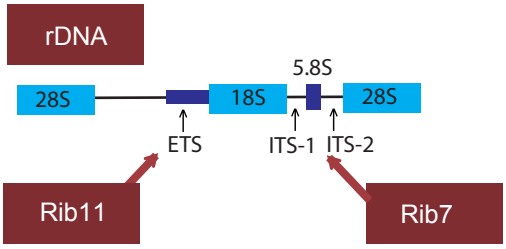
Southern blotting was repeated in a number of other haematological malignancies and hypomethylation was seen in a high proportion of cases ranging from 35% to 100% (based on 410 samples, Figure 3.1B).

3.2.3 DNA methylation in CLL at repetitive elements

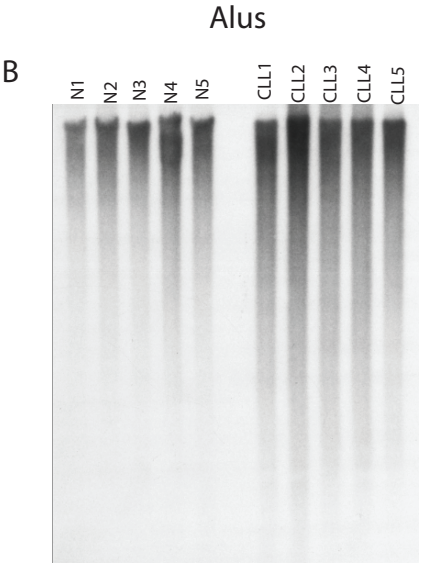
Prior to engaging in an extensive detailed comparison of methylation in normal and CLL samples, it was important to confirm that the CLL samples under study had expected methylation abnormalities, such as hypomethylation at repetitive elements.

Global DNA hypomethylation is a hallmark of malignancy and has been described in chronic lymphocytic leukaemia patients as assessed both by total 5mC content and overall susceptibility of genomic DNA to digestion by methylation-sensitive restriction enzymes (WAHLFORS *et al.* 1992). Global hypomethylation in CLL correlates with IGVH mutational status and thus prognosis, but not clinical stage (LYKO *et al.* 2004). Global hypomethylation has long been assumed to be due to hypomethylation at repetitive elements, as these make up a large percentage of the genome and are known to be densely methylated in normal tissues (WEISENBERGER *et al.* 2005). Although as discussed in the introduction, hypomethylation at repetitive elements has been reproducibly shown in a variety of other malignancies, it had not at the time of starting this study been directly investigated in CLL. Furthermore, investigation of the methylation at repetitive elements in other haematological malignancies has shown differential changes in methylation at interspersed and tandem repeats (CHOI *et al.* 2009). Figure 3.2A shows a schematic diagram of the different classes of repeats investigated here: LINE-1, *Alu*s (SINEs), alpha satellite repeats, subtelomeric repeats and ribosomal DNA arrays. By using digestion with the methylation-sensitive restriction enzyme HpaII and Southern blotting followed by quantification with a phosphorimager, the degree of loss of methylation at these repeats was assessed in the peripheral blood of five normal individuals and five individuals with CLL. Figure 3.2B shows the Southern blot of these samples hybridised with a probe for *Alu*s, and it can be seen that the smear created by the RE is longer in the CLL cases than the normals, indicating the presence of smaller

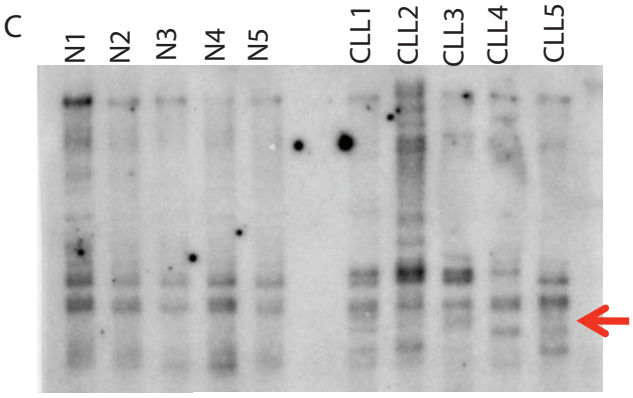




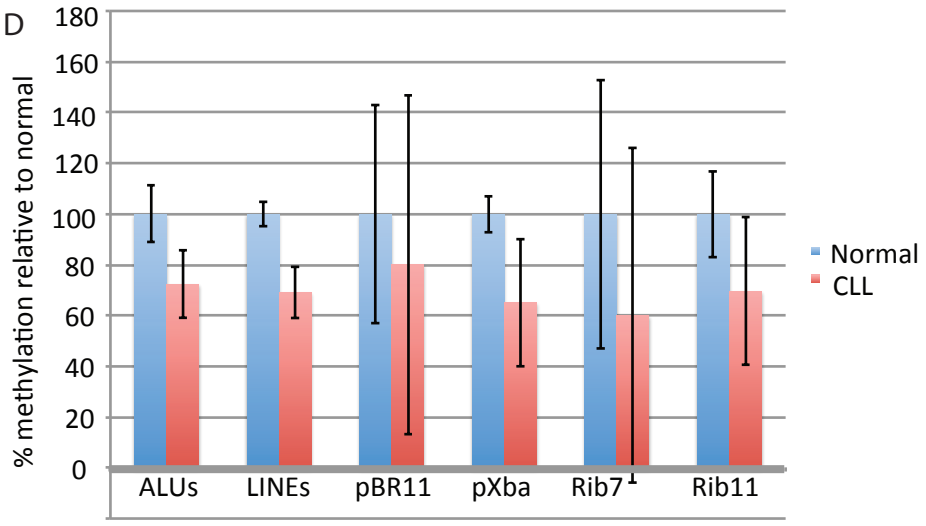
3.2A. Schematic diagram of different repetitive sequences in the human genome. Repeats tested by Southern blotting are highlighted in maroon.



3.2B. Southern blot of five normal and five CLL cases. The DNA was digested with BglII and HpaII, electrophoresed in an agarose gel, blotted to a membrane and hybridised with a probe corresponding to a sequence common to *Alu* repeats. The smear from CLL cases migrates further than for the normal cases, indicating smaller bands from more frequent cutting due to hypomethylation.



3.2C. Southern blot of the same CLL cases, hybridised with a probe corresponding to TelBam3.4. Dicrete bands of lower molecular weight in the CLL cases indicate hypomethylation (red arrow).



3.2D. Quantitation from phosphorimager. Each pair of bars represents a single Southern containing five normal cases (blue) and five CLL cases (red). The height of the bar represents the average quantitation for the five cases with standard deviations seen as error bars. The average quantitation is expressed relative to normal.

fragments, and thus hypomethylation. When this was quantitated, methylation was found to be decreased by 30% in both LINE-1 and *Alu* repeats as shown in Figure 3.2D. The hypomethylation was slightly more pronounced (35% reduction) in one alpha satellite repeat (pXba) but another (pBR11) showed no change in methylation between normal and CLL. A subtelomeric repeat (TelBam 3.4) also showed hypomethylation. Although it was not possible to quantify this, the CLL patients had smaller bands, indicating more frequent cutting by HpaII and therefore less methylation (Figure 3.2C). Two probes within the ribosomal DNA arrays were also tested (one within the Sal box upstream of the promoter (Rib11), and one in the 5.8S coding region (Rib7)), however only Rib11 seemed to indicate hypomethylation in CLL samples compared to normal, although to confirm this more samples would have needed to be tested.

In conclusion, similar to other haematopoietic malignancies, CLL is associated with hypomethylation, but this affects different classes of repetitive elements to a varying degree. Since carrying out this work, DNA methylation of repetitive elements in CLL has been reported in the literature, showing that *Alu*s, LINEs and alpha satellite repeats were less methylated in CLL compared to normal, and that in the case of alpha satellites, this was significantly associated with a poorer prognosis (FABRIS *et al.* 2011).

Once substantial methylation abnormalities were confirmed in CLL, rather than isolated at alpha globin, the next stage was to examine the pattern of CpG methylation on a genome-wide basis using the MeDIP (methylated DNA immunoprecipitation) technique, which had recently been published in Nature and seemed promising (WEBER *et al.* 2005). However, this needed to be established in our lab, and rigorous testing would require methylation data at individual CpGs in a test area. To this end, bisulfite sequencing was used to determine the CpG methylation status of 330 CpGs in chromosome 16p at the alpha globin locus and surrounding region while in parallel establishing the MeDIP technique. In the next section I will present the findings related to the bisulfite sequencing results.

3.3 Conventional Bisulfite Sequencing (CBS)

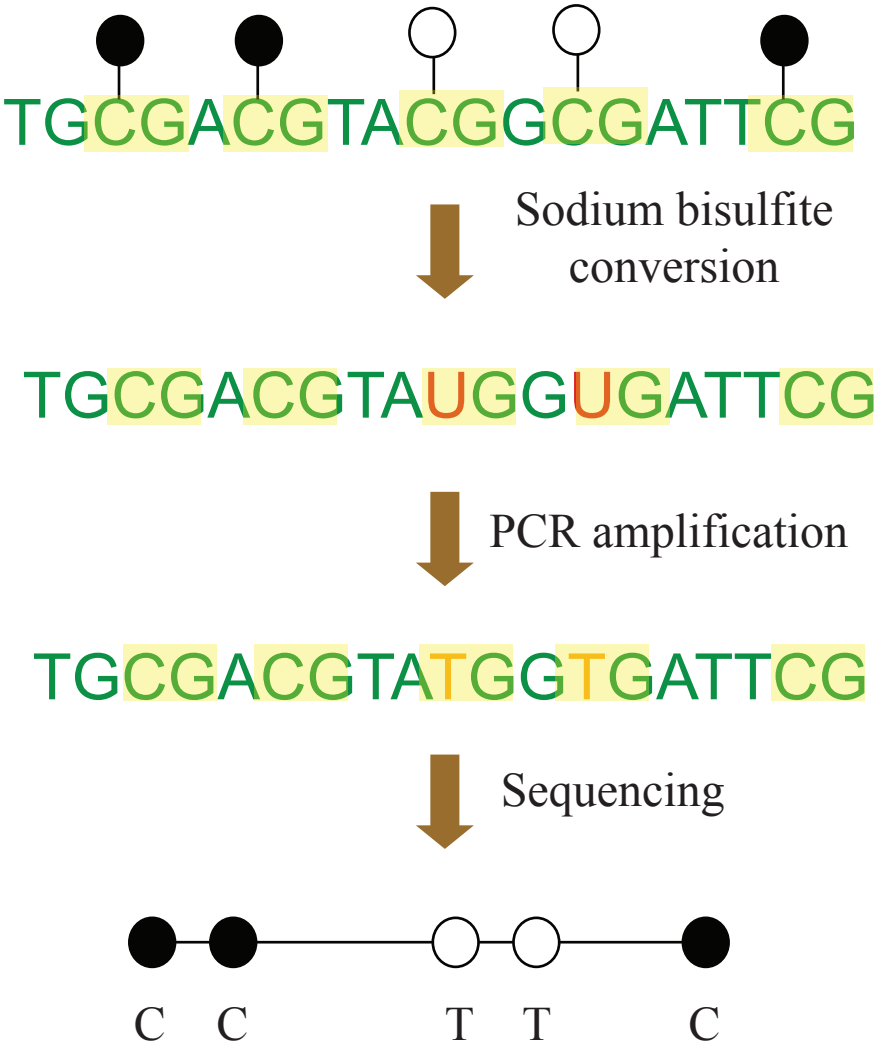
3.3.1 Overview of the technique

This technique, often referred to as the gold standard (FROMMER *et al.* 1992) for methylation studies, is based on the differential conversion of methylated (5mC) and unmethylated (C) cytosines on exposure to sodium bisulfite (Figure 3.3). The reaction converts unmethylated C to uracils, but 5mC are protected and remain as cytosines. Following PCR amplification and sequencing, thymines are incorporated in place of uracils. Therefore, the sequencing reaction will yield CG for methylated CpGs and TG for unmethylated CpGs in the original sequence.

3.3.2 Technical considerations

The region covered spans 128kb across 16p, of which 12.4kb were bisulfite sequenced. The amplicons sequenced are summarised on Figure 3.4. DNA extracted from whole blood was converted using sodium bisulfite and then amplified by nested PCR using primers specific to modified DNA. The efficiency of conversion was 99.5%, based on the assessment of 3167 non-CpG cytosines. Commercially available fully methylated and fully unmethylated DNA (Millipore) were used as positive and negative controls. Methylation levels were assessed semi-quantitatively. The proportion of methylated cytosines was estimated for each CpG from the heights of the C vs T peaks in the sequencing reaction. This is illustrated in Figure 3.5A and shows how CpGs were scored according to the relative C vs T peak heights. Where the C peak was smaller than the T peak, the CpG was scored as "25%" although true methylation could be anything between 0 and 50%. Similarly for methylations between 50-100%, the CpGs were scored as "75%". In order to assess the accuracy of this semi-quantitative method, subcloning and sequencing of several clones was carried out for regions of high, medium and low methylation levels. As illustrated in Figure 3.5B, there was excellent concordance of the results between subcloning and the semi-quantitative method, for a variety of methylation levels: Me15 has high methylation throughout the amplicon, Me30 CLL has low methylation with two highly methylated CpGs, and Me30 normal has intermediate methylation. Subcloning was carried out for eight separate loci, each showing good concordance.

Figure 3.3. Bisulfite sequencing. Unmethylated cytosines (white lollipops) are converted to uracils on exposure to sodium bisulfite while methylated cytosines (black lollipops) are protected from conversion. During PCR amplification, uracils are converted to thymines. In the final sequencing reaction, CG dinucleotides indicate CpGs which were methylated in the original template DNA, while TG dinucleotides represent unmethylated CpGs.



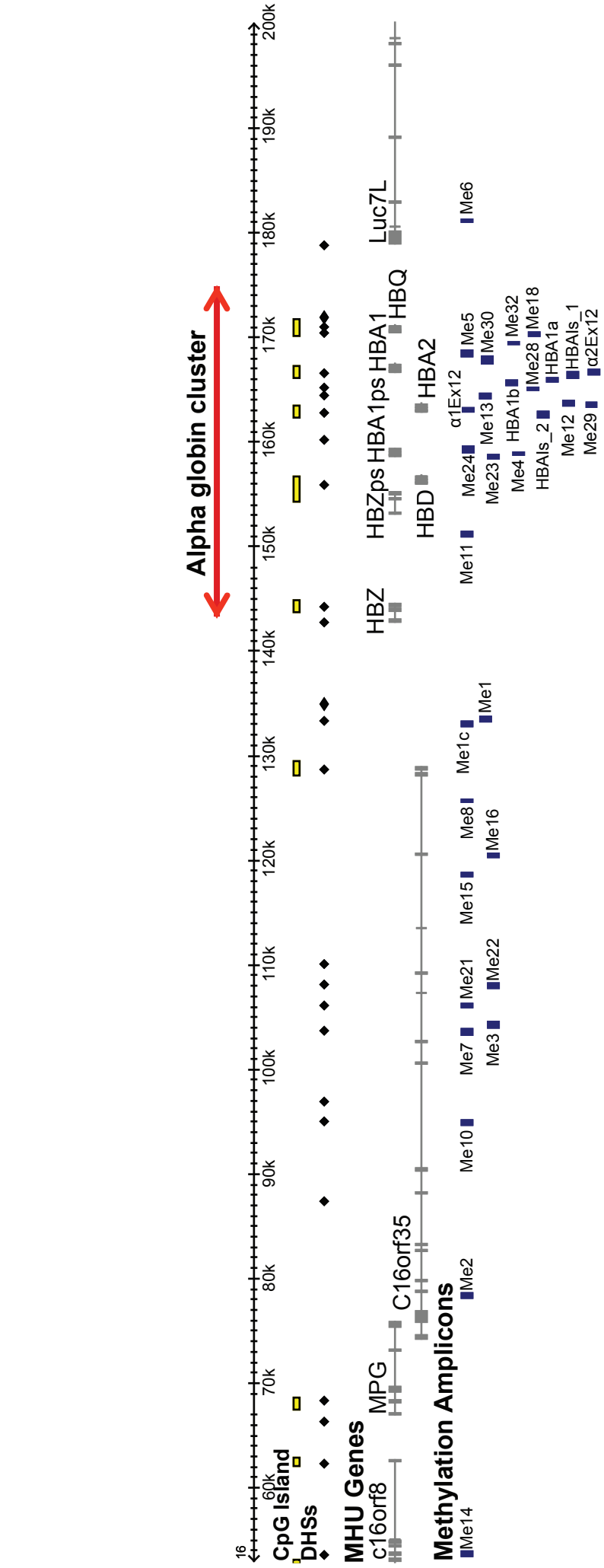


Figure 3.4. Map of the 16p region showing bisulfite sequencing amplicons, visualised on in-house 16p browser.

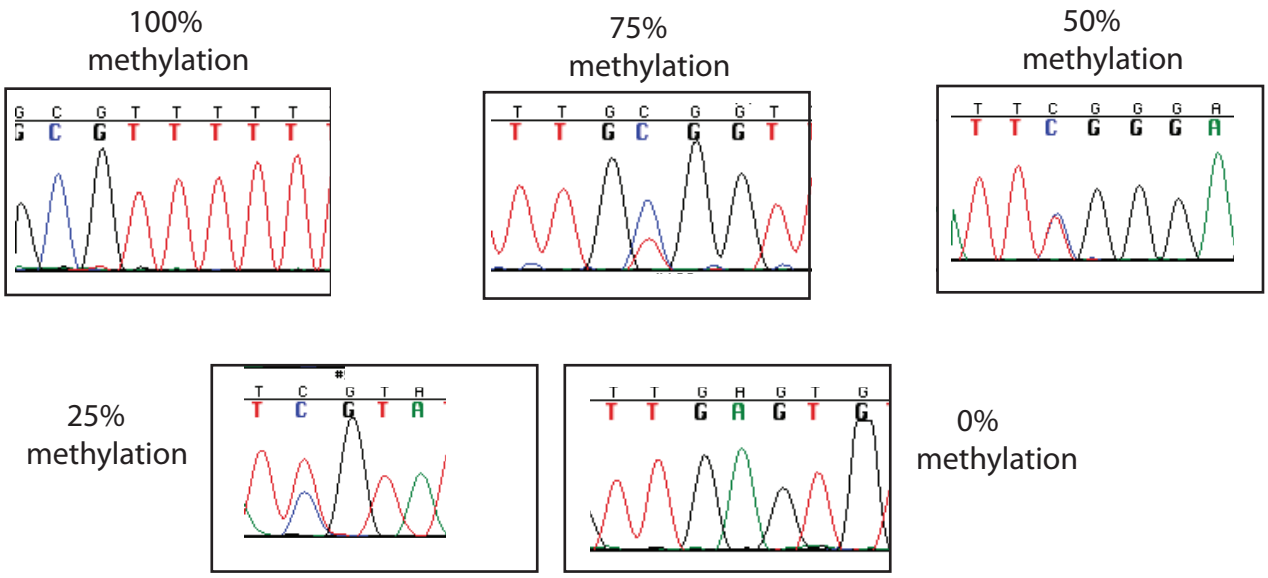


Figure 3.5A. Example peaks from bisulfite sequencing demonstrating how the semi-quantitative methylation calls were made. Peaks that were completely C were called 100% methylation, those where C and T peaks were roughly equal were called 50%, those which were completely T were called 0%. Positions where the C peak was higher than T were called 75% and those where T were higher than C were called 25% methylated.

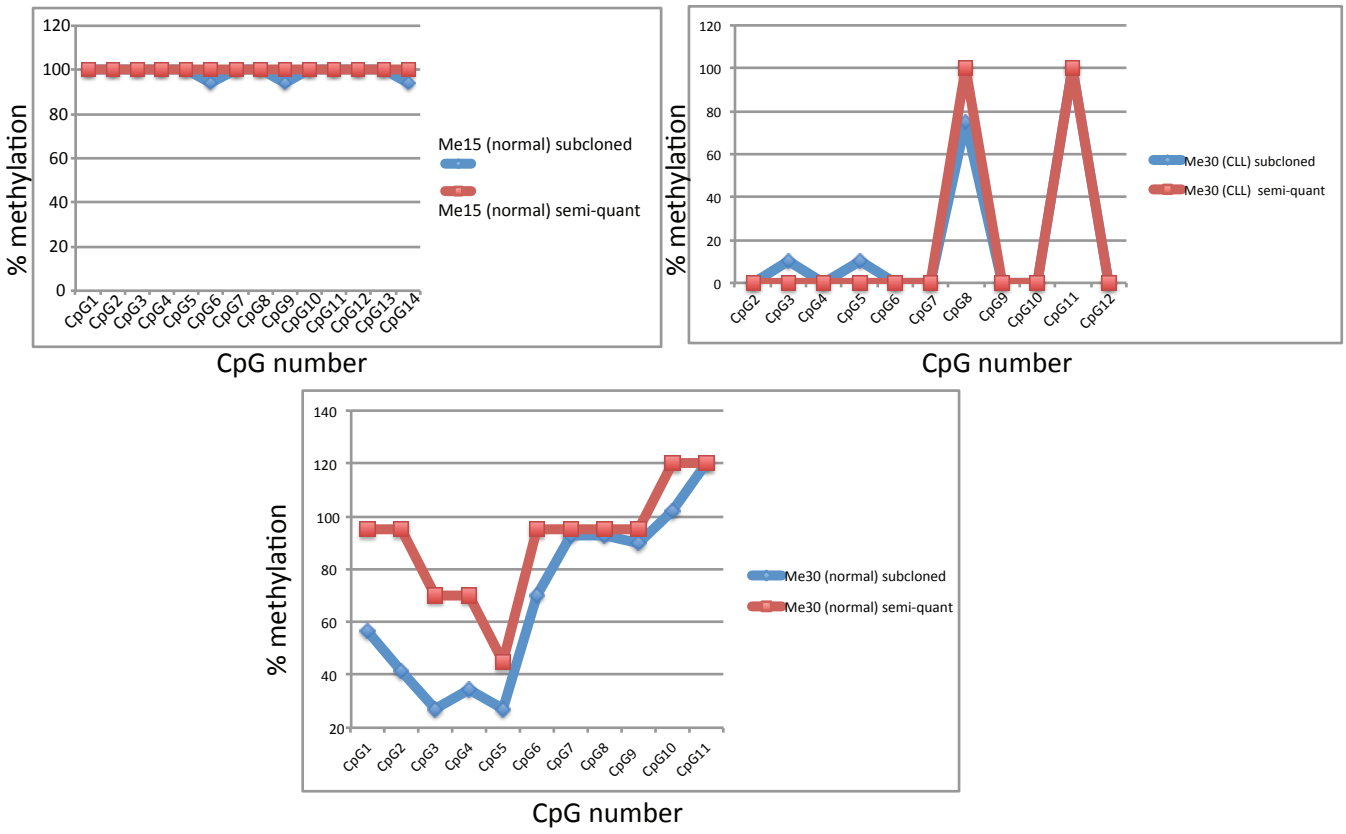


Figure 3.5B. Validation of the semi-quantitative technique. Samples were sequenced and the % methylation at each CpG determined semiquantitatively ("semi-quant" curves). The PCR products were then subcloned into DH5α cells and individual clones were sequenced. The average of the total number of clones is then shown on the "subcloned" curve. Number of clones sequenced were 18 for Me15(normal), 14 for Me30(normal) and 12 for Me30(CLL).

Figure 3.6 shows an example amplicon. Me11 is an amplicon in an intergenic region, between the genes *HBZ* (zeta globin) and *HBZps* (pseudozeta globin) (chr 16: 150917-153129, see Figure 3.4), and contains 12 CpGs. The CLL and normal samples are all fully methylated at CpGs 1, 2, 11 and 12, and fully unmethylated at CpG7. CpG3 shows hypomethylation in both CLLs compared to normal (Figure 3.6B), as does CpG10. CpGs 4, 6, 8 and 9 show hypomethylation in one of the CLL samples only. CpG5 shows discordance between the two samples and is scored "mixed".). The legend in Figure 3.6 explains how DNA methylation levels will be visualised in this thesis.

3.3.3 Choice of samples

3.3.3.1 CLL samples

Patient samples

The CLL samples were obtained from untreated patients prior to entering a therapeutic clinical trial. CLL was diagnosed based on morphological assessment of peripheral blood, following which immunophenotyping was carried out. White cell counts varied between 23.5-159x10^9/L. The cases were all CD5 positive, CD20 positive, sIg positive, CD23 positive, FMC7 negative and had variable expression of CD79b. The cells were isolated by Ficoll gradient with typical CD19 counts of 80-88% on immunophenotyping.

CLL cell lines

For some of the experiments listed below, it was not possible to use fresh CLL samples due to inadequate amounts of DNA or protein available and therefore CLL cell lines had to be used. These were commercially obtained from DSMZ and all three represent CLL/B-PLL patient samples that were immortalised by EBV (Ebstein-Barr virus). This does not constitute ideal experimental conditions, as cell lines are known to contain a greater amount of 5mC compared to equivalent fresh tissues (FLATAU *et al.* 1983), and in particular to have hypermethylated CpG islands (SMIRAGLIA *et al.* 2001) (PAZ *et al.* 2003). Nevertheless, from necessity the use of cell lines is common in DNA methylation studies.

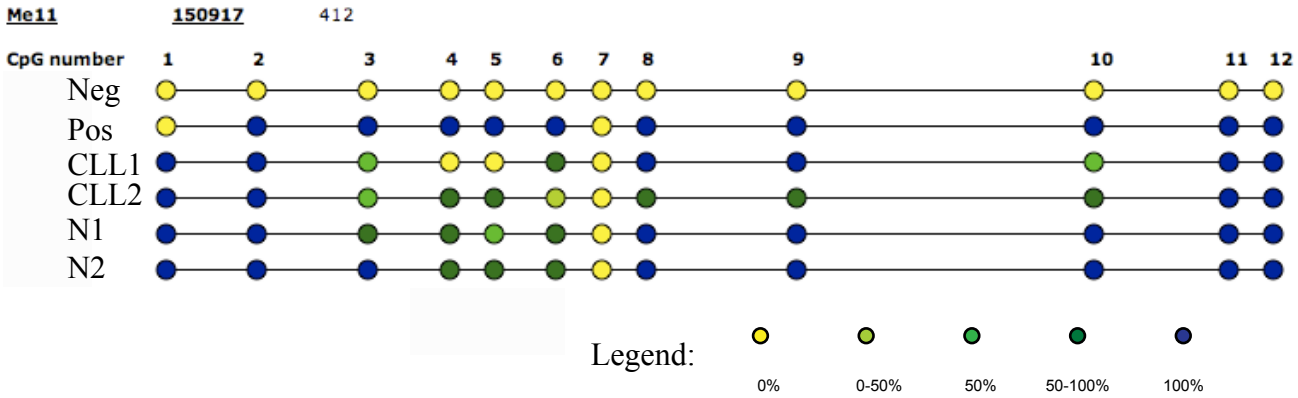


Figure 3.6A Me11 amplicon. There is hypomethylation in the two CLL samples at CpGs 3 and 10. There is hypomethylation in one of the CLL samples at CpGs 4, 6, 8, and 9. CpG5 is "mixed". CpG 7 is completely unmethylated in all samples, while CpGs 1, 2, 11 and 12 are fully methylated. Note the positive control has unmethylated CpGs 1 and 7, reflecting the imperfect nature of *in vitro* methylated samples. Controls were obtained from Millipore (CpG Universal Methylated and Unmethylated Genome DNA).

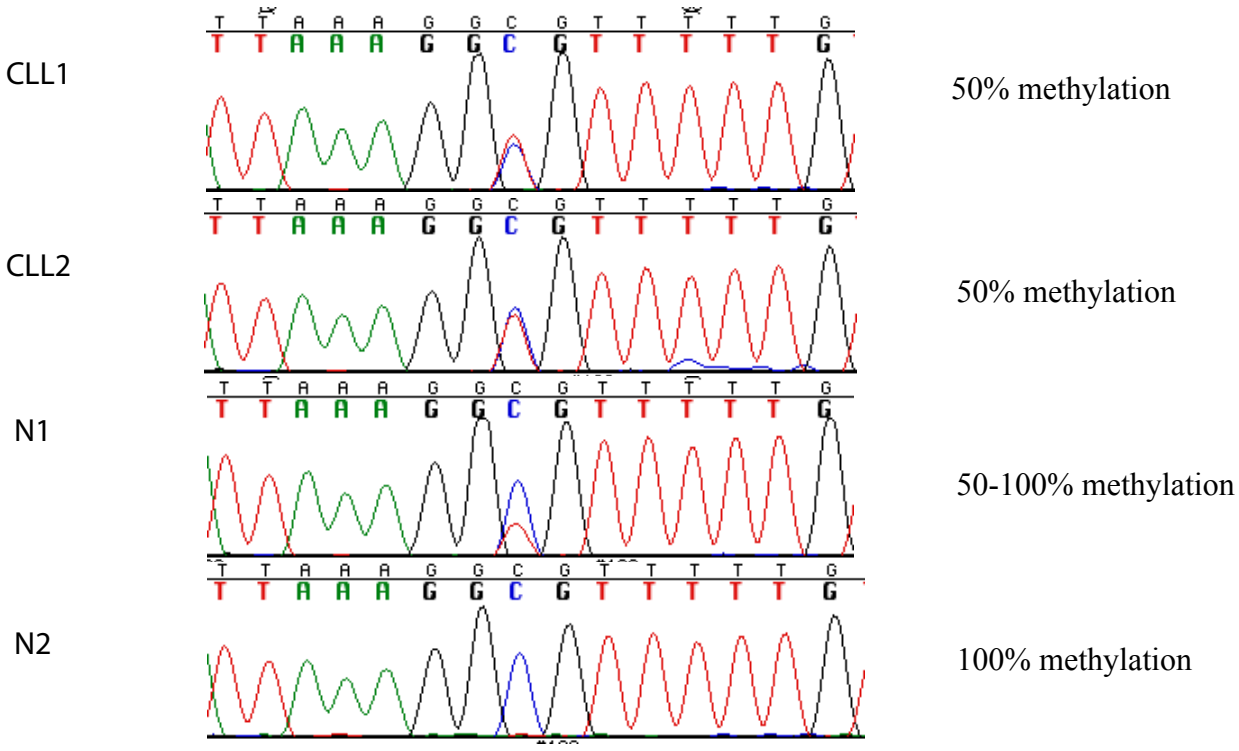


Figure 3.6B Me11 CpG3. Sequencing result for Me11 CpG3 showing 50% methylation in CLL1 and CLL2 (T and C peaks of equal heights), 100% methylation for N2, and methylation between 50 and 100% for N1, called "75% methylation.

3.3.3.2 Normal samples

The normal counterpart to CLL

The key question prior to embarking on an assessment of the differences in methylation between normal and CLL relates to which normal cell type should be used for comparison. This raises the issue of the normal counterpart to the CLL B-cell, which is one of longstanding debate (BERTILACCIO *et al.* 2010). It was long thought that the malignant cells in the peripheral blood of CLL patients were terminally differentiated non-dividing cells accumulating due to a defect in apoptosis (DAMESHEK 1967). Contrary to this idea, recent studies using *in vivo* labelling of CLL cells with heavy water (MESSMER *et al.* 2005), surface immunophenotyping (SCHROEDER and DIGHIERO 1994) and telomere lengths (DAMLE *et al.* 2002) show that this is likely not the case, and that CLL cells in peripheral blood are actively dividing, at a rate between 0.1-1% of the entire clone daily. Based on gene expression profiling and surface membrane immunophenotype, the normal counterpart to the CLL cell seems to be an antigen experienced B-cell (KLEIN *et al.* 2001). As clinically there are two very different CLL entities based on IGVH mutational status, each associated with quite a different prognosis, the possibility has been raised that there may be at least two precursor cells (CHIORAZZI and FERRARINI 2011). Currently the generally accepted view is that CLL arises from marginal zone B-cells, which can respond to bacterial polysaccharides in a T-cell independent manner, especially as these cells can have a mutated or unmutated IGVH (CHIORAZZI and FERRARINI 2011). However, one of the hallmarks of CLL cells is the fact that they are CD5 positive, whereas marginal zone B-cells are CD5 negative. It is thought that CD5 may become positive during the process of immortalisation, as it has been shown that CD5 may become upregulated with activation (KAPLAN *et al.* 2001).

EBV-immortalised B-cells

Due to the lack of certainty regarding the normal counterpart to CLL, and the absence of CD5 positive B-cells in normal peripheral blood, the best cell to use for comparison was thought to be the CD19 positive B-cell. B-cells make up 20% of lymphocytes, which themselves make up 40% of leucocytes. While in theory from 30ml of whole blood with an average leucocyte count of $5x10^6$/ml, this

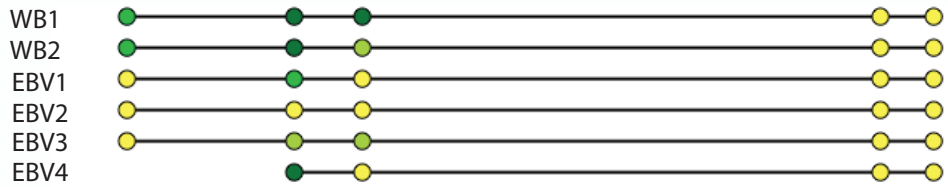
should yield 1.2x10^7 B-cells, in reality yields were much lower (see *"CD19 cells"* below). Due to their paucity in peripheral blood, the use of EBV-immortalised B-cells in cell culture was therefore considered. However, a detailed assessment of 2.4kb in chromosome 16p by bisulfite sequencing revealed their methylation to be abnormal compared to CD19 B-cells from peripheral blood. It has previously been reported that EBV-immortalisation is associated with DNA methylation abnormalities (SUN *et al.* 2010). When comparing methylation levels between CD19 cells and EBV-immortalised lymphocytes, variability was seen as illustrated on Figure 3.7. This shows two amplicons, Me13 (Figure 3.7A) where CpGs 1-3 are hypomethylated in EBVs compared to whole blood and Me21 (Figure 3.7B) with variable methylation of CpGs 1-5. This therefore excluded the possibility of using these as CD19 substitutes.

CD19 cells

CD19 cells were obtained initially by MACS (magnetic-activated cell sorting) separation of whole blood, and later by Dynabead separation from lymphocyte cones. Whole blood was centrifuged using a histopaque gradient, which yields a mononuclear cell fraction, which was subjected to MACS separation, and an erythrocyte/granulocyte pellet from which neutrophils are readily isolated following red cell lysis. From 30-100ml of WB, 1.2x10^6 to 4x10^7 CD19 cells were obtained using MACS separation. Figure 3.8A shows cytospins prepared from different leucocyte fractions and Figure 3.8B shows the purity of these from FACS (Fluorescence-activated cell sorting) analysis. While the purity for CD3 and CD14 cells was 80-85%, it was much lower (60%) for CD19 cells. This seemed to be in part due to contamination from CD14 cells (11% of the population). To improve upon this, CD14 cells were removed prior to selecting with CD19. Although, where possible, CD19 cells were used, for some experiments (e.g. MeDIP) needing larger quantities of DNA, whole blood leucocytes were used. This therefore required confirmation that the methylation status of whole blood leucocytes and its different white cell fractions was the same or similar enough. This is shown in Figure 3.9. Panel A shows a Southern blot with BglII/HpaII digests and a probe for L0/L1 (HS-10) for whole blood and

A. Me13

CpG number	1	2	3	4	5
WB1	50	75	75	0	0
WB2	50	75	25	0	0
EBV1	0	50	0	0	0
EBV2	0	0	0	0	0
EBV3	0	25	25	0	0
EBV4		50	0	0	0



B. Me21

CpG number	2	3	4	5	6	7
WB1	75	75	25	100	100	100
WB2	100	50	25	100	100	100
EBV1	75	0	0	50	100	100
EBV2	75	75	25	75	100	100
EBV3	75	25	0	25	100	100
EBV4	75	25	50	100	100	100

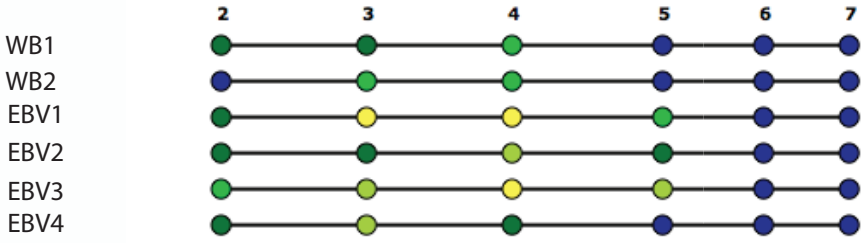


Figure 3.7A,B Two amplicons, Me13 and Me21, comparing the methylation status of EBV-transformed lymphocytes (EBV) with that of whole blood leucocytes (WB). The EBV lymphocytes are hypomethylated at CpGs 1-3 in Me13 whereas in Me21 EBV1-3 are hypomethylated at CpG4 while EBV4 is hypermethylated. This constitutes too great a variability to warrant use of EBVs as "normal lymphocyte" susbstitutes. For each amplicon, the top panel shows methylation % and the bottom panel shows the colour-coded scale.

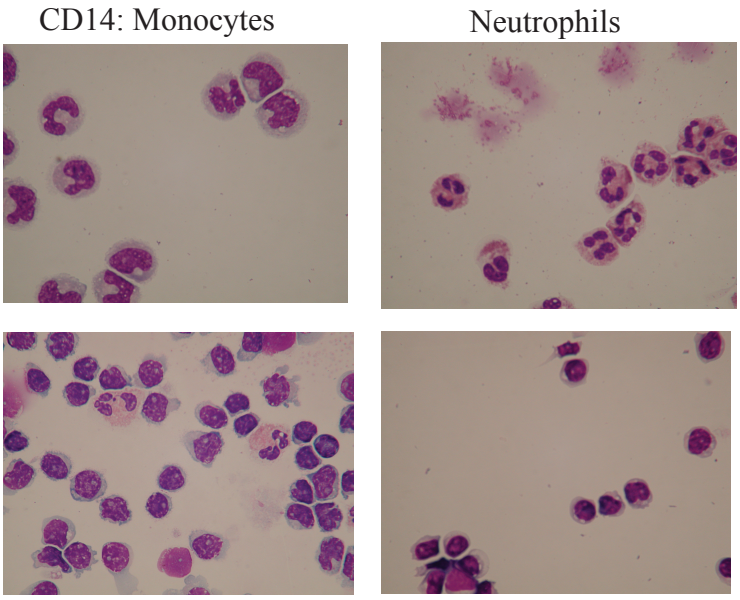


Figure 3.8A. Cytospins of the different white cell fractions as separated by histopaque and MACS

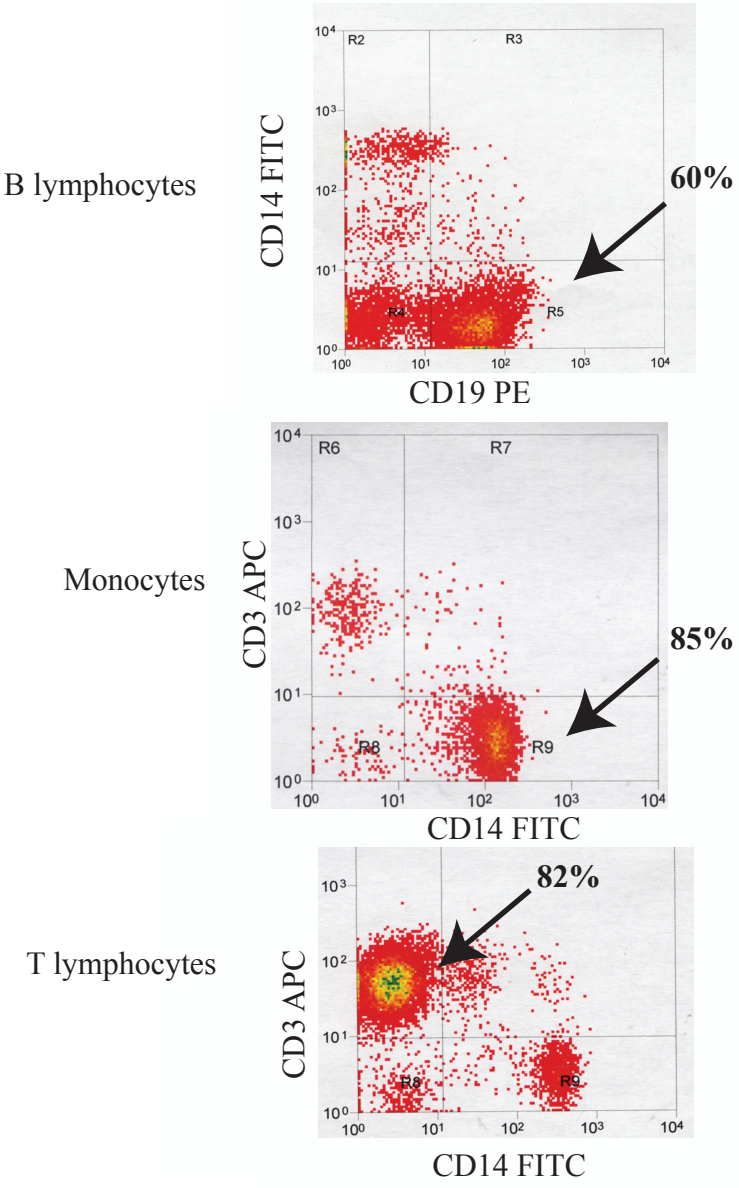
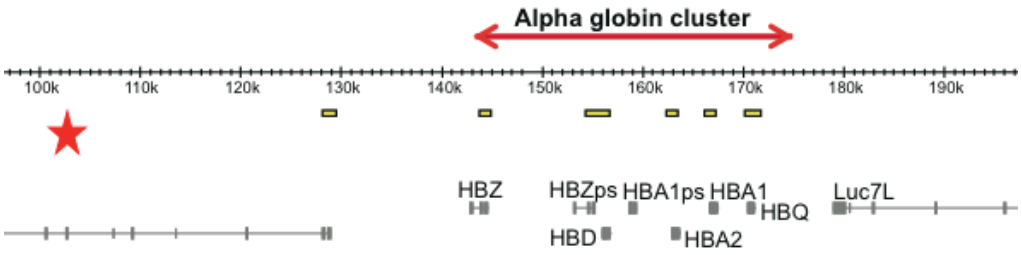
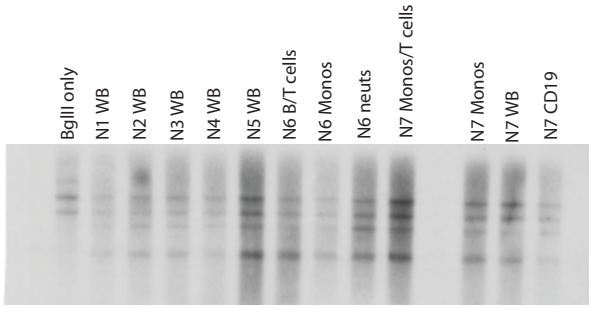


Figure 3.8B. FACS plots of the different white cell fractions. The CD19 fraction was the least pure (60%), due to contamination with CD14 monocytes (11%).

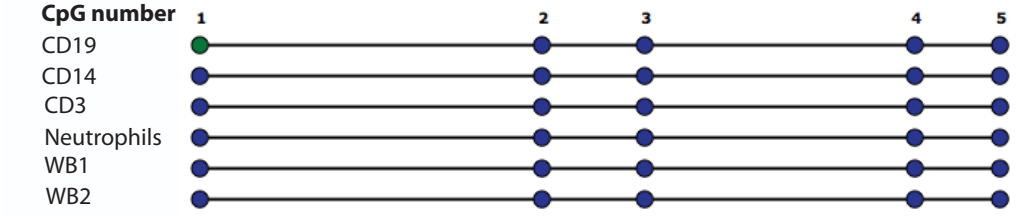


A



B

CpG number	1	2	3	4	5
CD19	75	100	100	100	100
CD14	100	100	100	100	100
CD3	100	100	100	100	100
Neutrophils	100	100	100	100	100
WB1	100	100	100	100	100
WB2	100	100	100	100	100



C

CpG number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
CD19	75	75	100	75	50	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD14	75	75	100	75	75	50	50	25	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD3	75	75	75	75	75	75	50	50	50	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Neutrophils	75	75	75	75	75	50	50	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WB1	75	75	75	75	50	50	50	25	0	0	0	0	25	25	25	25	25	0	0	0	0	0	0	0	0	0	0
WB2	75	75	100	75	50	50	25	0	25	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

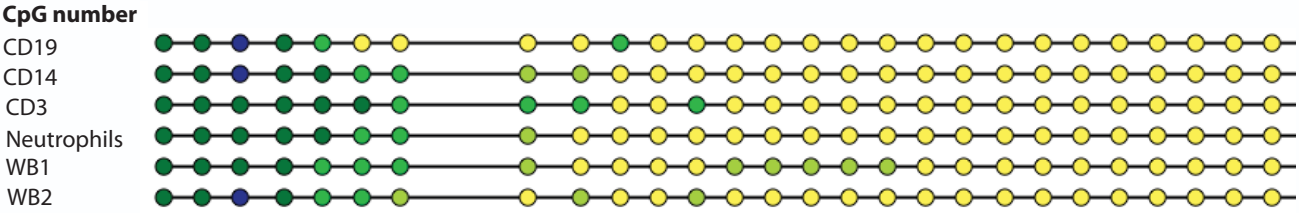


Figure 3.9 A. Southern blot with BglII/HpaII digests and a probe for L0/L1 (HS-10) for whole blood and different cell types which shows concordance of methylation between white cell types. Star indicates location of probe. Marker did not come out on this blot, but size estimations are by comparison with similar gel. **B.** Individual CpG methylation by bisulfite sequencing at Me3, showing very similar methylation levels between white cell types in a region of high methylation. **C.** Further bisulfite sequencing at Me18, showing very similar methylation levels between white cell types in a region of medium to low methylation.

different cell types and shows concordance in the methylation levels between white cell types. This is in agreement with methylation studies across cell types that showed <5% differences in methylation between CD4 and CD8 T-cells (ECKHARDT *et al.* 2006b). Panels B and C illustrate individual CpG methylation for two amplicons and show very similar methylation levels at individual CpGs in whole blood compared to lymphocytes.

3.3.4 Changes in DNA methylation in CLL

3.3.4.1 Technical details

Initial bisulfite sequencing was carried out on two CLL samples and two normal samples. Whole blood was used, as these were the samples with which MeDIP would be carried out, and using identical samples for both experiments would ensure a direct comparison could be made. The amplicons sequenced cover 391 CpGs, of which 330 were informative. The remainder were not present on either the forward or reverse sequencing reactions due to technical limitations. Each of the 330 informative CpGs was assessed for its status in CLL compared to normal as to whether methylation increased, decreased or stayed the same in CLL samples vs normal. Where there was discordance between the two CLLs or the two normals, this was scored as "mixed".

3.3.4.2 Direction and magnitude of changes from normal to CLL

Of the informative CpGs, the majority (62%) was unchanged between normal and CLL samples. Twelve percent of CpGs assessed were hypermethylated in CLL over normal and 19% hypomethylated in CLL. Hyper- and hypomethylation were determined if there was >20% change. Four percent were mixed and three percent uninterpretable (Figure 3.10A).

3.3.4.3 Distribution of changes

To look for patterns in the methylation changes in CLL, I assessed whether the CpGs that changed methylation status occurred in particular regions- introns, exons, repetitive sequences, DNA hypersensitive sites, CpG islands (WILKIE *et al.*), or intergenic regions. Figure 3.10C shows the distribution of the CpGs assessed, based on the change in methylation in CLL. It can be seen that CpGs that become hypomethylated tend to be in intergenic regions, while more of the CpGs that

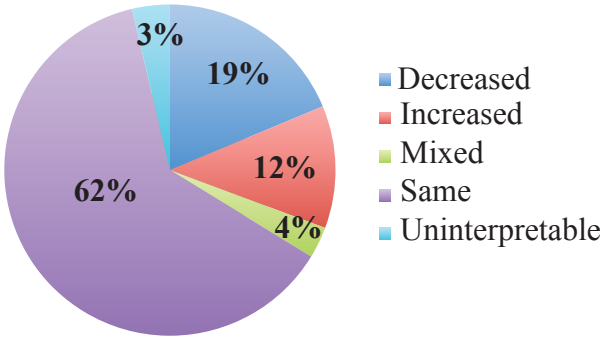


Figure 3.10A Distribution of CpG changes in CLL compared to normal

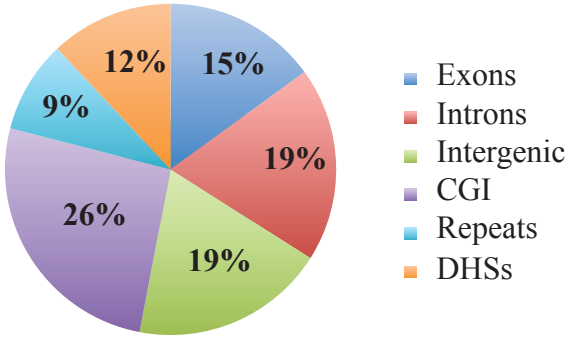


Figure 3.10B Distribution of CpGs per genomic feature in 16p region

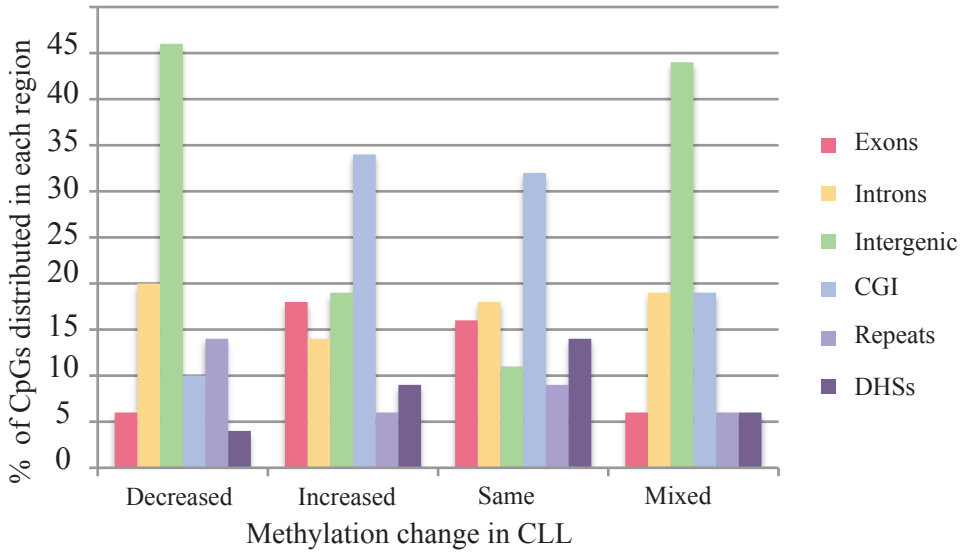


Figure 3.10C Distribution of CpG per genomic feature as a function of changes in CLL compared to normal

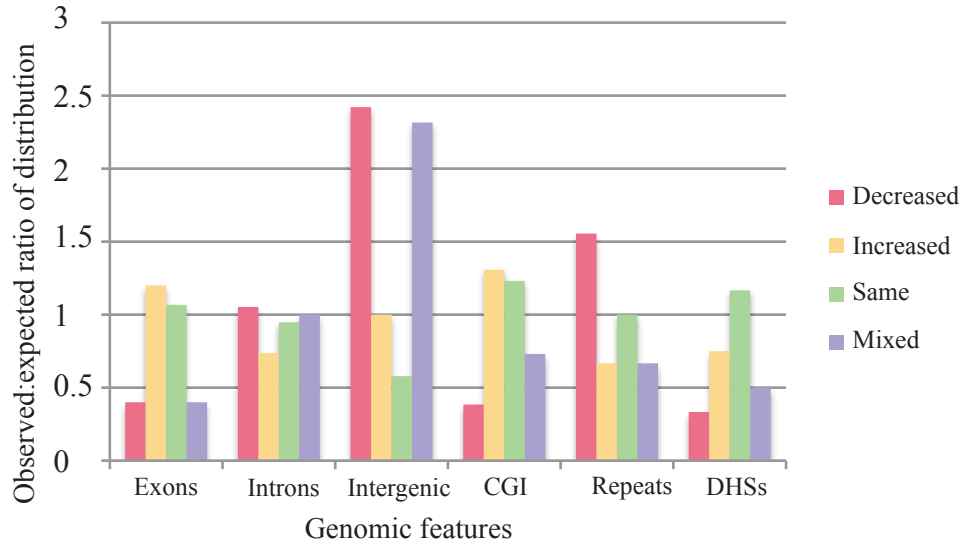


Figure 3.10D Observed:expected ratio of CpG changes in CLL compared to normal as a function of distribution of CpGs amongst genomic features.

became hypermethylated were in CGIs than in other regions. The CpGs that were unchanged also were more frequently in CGIs. However it is not possible to say, for example, whether the presence of a CpG in an intergenic region increases its likelihood of becoming hypomethylated in CLL as would be suggested by this graph, without taking into consideration how the CpGs assessed are distributed amongst these genomic regions. These are distributed 15% in exons, 19% in introns, 19% in intergenic regions, 26% in CGIs, 9% in repeats and 12% at DHSs (Figure 3.10B). It is now possible to calculate an observed:expected (O:E) ratio of the distribution of each set of CpGs (hypermethylated in CLL, hypomethylated, same or mixed). This is shown in Figure 3.10D. It is therefore clear that the presence of a CpG in an exon decreases its likelihood of becoming hypomethylated in CLL (O:E ratio 0.4). Their presence in an intron, however, does not affect their likelihood of change, but their presence in intergenic regions increases their chance of becoming hypomethylated (O:E ratio 2.4).

CpGs in CGIs are less likely to become hypomethylated, which is logical as these are already unmethylated in normal samples. CpGs in repeats, however, are more likely to become hypomethylated in CLL. These data are consistent with previously published findings that repeats become hypomethylated in most malignancies, as discussed in Chapter 1.

3.3.4.4 Motif analysis

Motif analysis was carried out to determine whether the regions where changes in DNA methylation are observed can be predicted by sequence motifs. Using the MEME software suite, the 7 most hypermethylated CpGs (>2SD from the mean) were searched for common motifs. These are illustrated on Figure 3.11A. Motif 1 was [GA]CG[CT][GC]C[GTA]G[TG][CT]C and was found in 5 of the 7 sequences representing hypomethylated CpGs, and 9 of the sequences representing all the CpGs, giving an observed to expected (O:E) ratio of 24. Motif 2, [CT]T[CT]A[AT]T, was found in 2 of 7 hypermethylated CpGs sequences with an O:E ratio of 47, and Motif 3, CGGGAT, also found in 2 of the 7 hypermethylated CpG sequences but with an O:E ratio of 10.

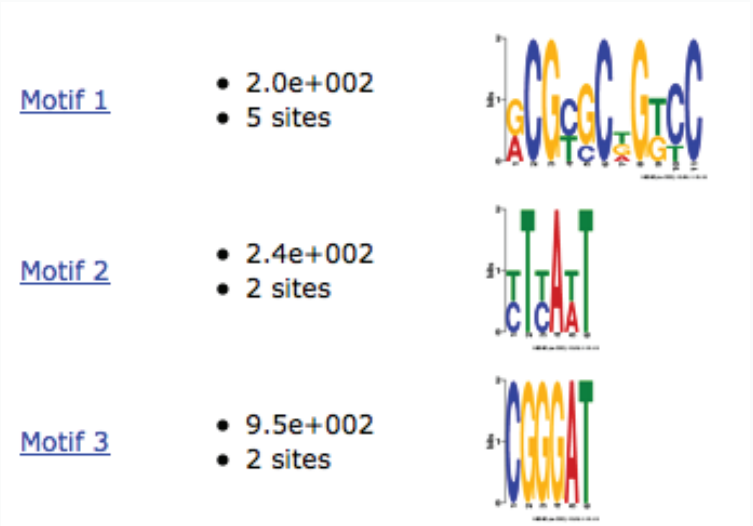


Figure 3.11A Motifs associated with CpGs hypermethylated in CLL

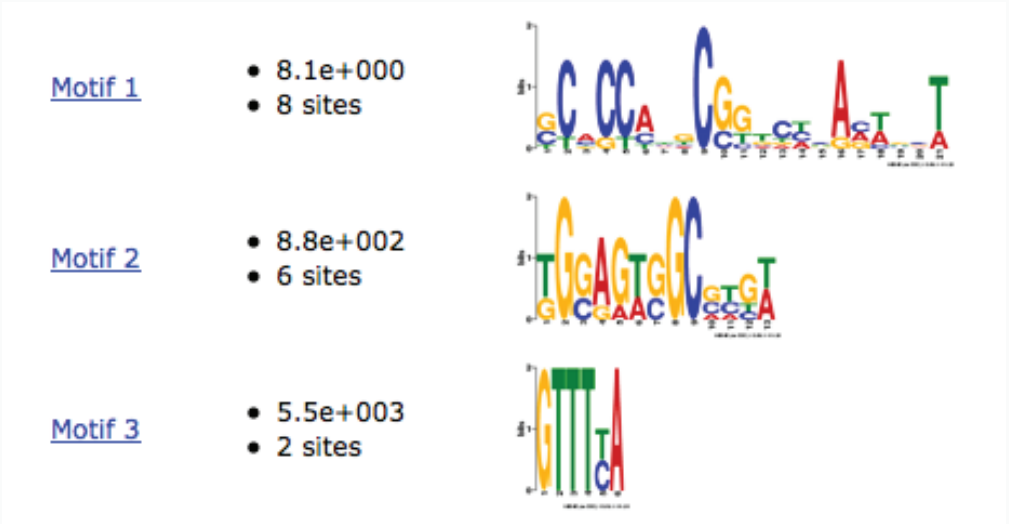


Figure 3.11B Motifs associated with CpGs hypomethylated in CLL

The same analysis was carried out for the 21 most hypomethylated CpGs, and is shown in Figure 3.11B. Motif 1 was too long to be of interest. Motif 2, [TG]G[GC]AG[TA][GC]GC[GC][TC]G[AT] was found in 6 of 21 hypomethylated CpG sequences and not at all in the CpGs not hypomethylated and Motif 3 GTT[TC]A was found in 2 of the 21 hypomethylated CpG sequences with an O:E ratio of 3.

Although this analysis did not identify a unique over-represented motif, it was informative, showing that some DNA sequences may be over-represented at sites of both hyper- and hypomethylation in CLL.

3.3.4.5 Extension to further samples

As this work was carried out on only two normal and two CLL samples, it was important to determine whether these changes in methylation were reproducible in other samples. To this end, the bisulfite sequencing analysis was extended to 14 normal samples and 14 CLL cases, for a total of 6 amplicons where there was significant hypomethylation identified in the initial analysis. These results are shown in Figure 3.12 and show the overall methylation at individual CpGs for the average of the 14 normal and the 14 CLL cases. As can be seen, the differences picked up by CBS on 2 normal and 2 CLL cases are reproducible across larger sets of samples, and the hypomethylation is statistically significant at 21 of the CpGs studied. These "consistently hypomethylated CpGs" form the basis of further work in the thesis (see Chapter 5).

With good quality data for methylation abnormalities in 16p, it was now possible to implement MeDIP and validate it against this dataset.

3.4 Methylated DNA immunoprecipitation (MeDIP)

3.4.1 Method overview and rationale

MeDIP is based on the principle of immunoprecipitating DNA using a monoclonal antibody against methylated cytosines, amplifying the DNA, labelling it and hybridising it to a microarray (in this case, an ENCODE array, covering 1% of the genome) (Figure 3.13). In theory, this should provide a global unbiased

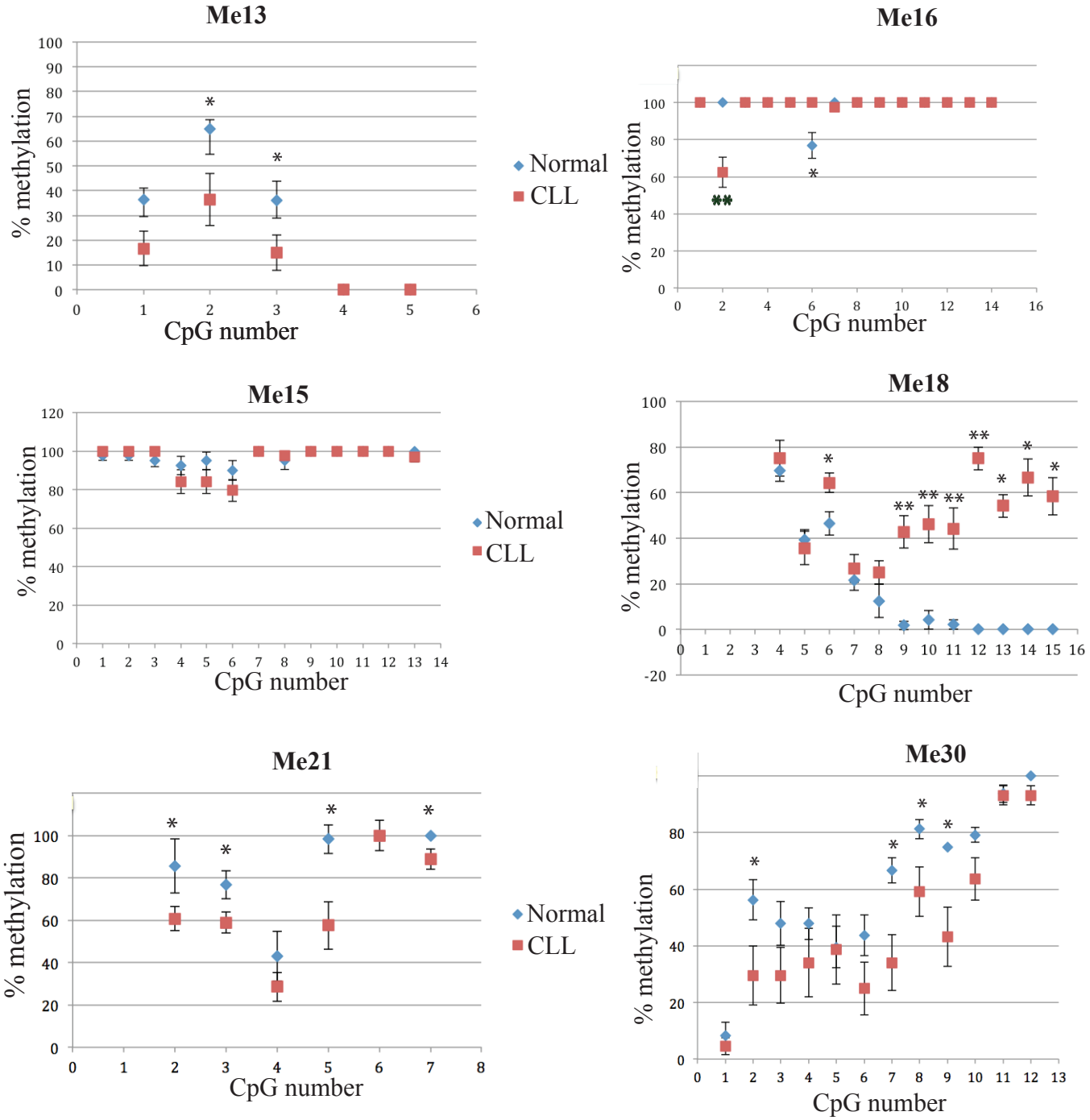


Figure 3.12 Overall methylation at individual CpGs in six different amplicons for the average of 14 normal samples (blue diamonds) and 14 CLL samples (red squares). Differences that were statistically significant as assessed by Student's t-test are starred.

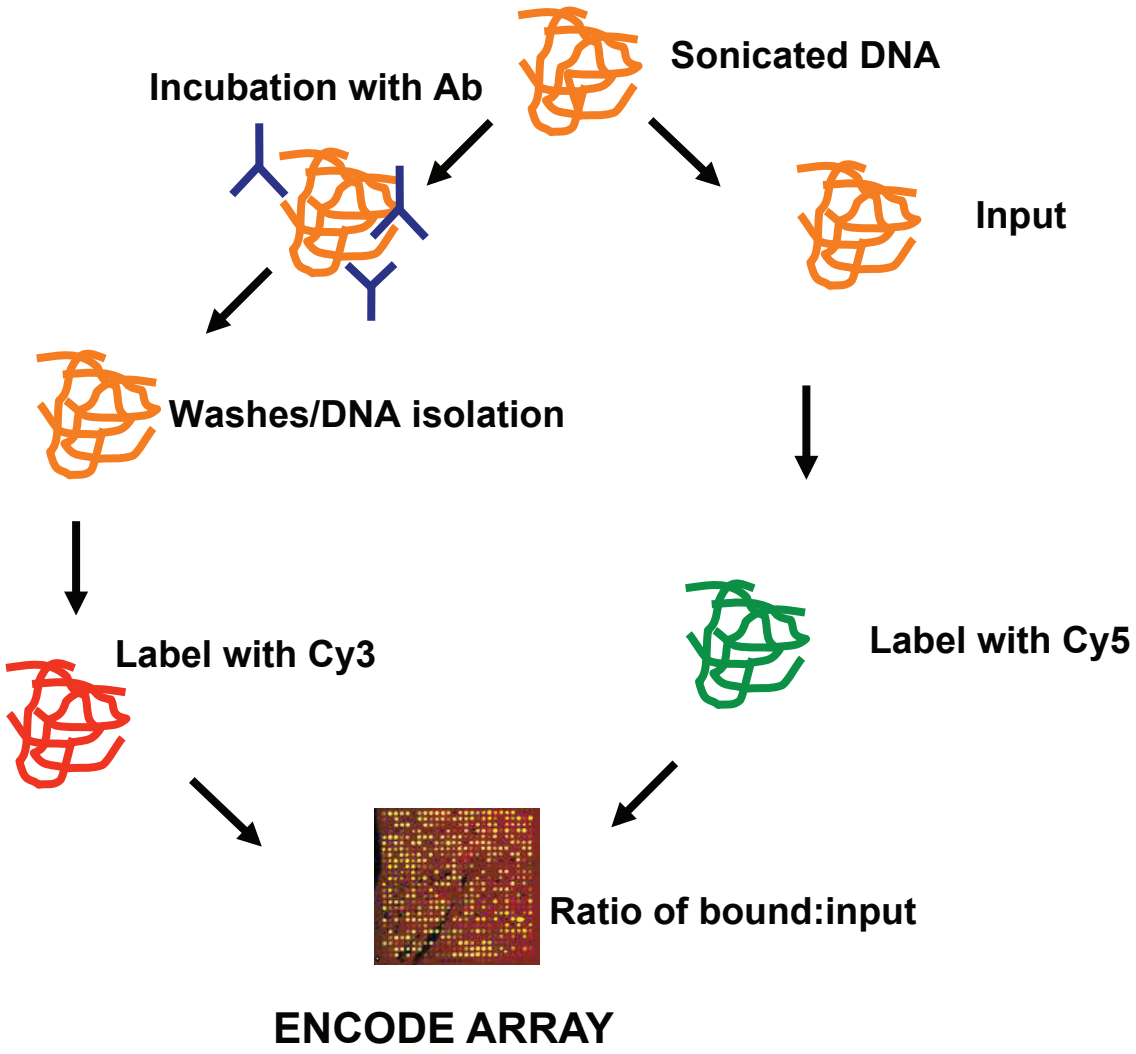


Figure 3.13 Principle of MeDIP reaction. Sonicated genomic DNA is incubated with both a monocloncal antibody to 5-methylcytosine and beads. Following washes, the precipitated DNA is isolated and, after amplification, labeled with one dye while input DNA is labeled with another. Hybridisation to a microarray with both fractions demonstrates which DNA sequences were immunoprecipitated, and therefore methylated.

assessment of the methylation status of all CpGs, albeit not at single CpG resolution. The starting material is sonicated DNA, therefore MeDIP will assess methylation over a range of CpGs present in various fragments. Furthermore, this assessment should be quantitative, as the amount of bound:input signal generated should be proportional to the amount of DNA precipitated by the antibody, itself a function of the proportion of methylated CpGs. The initial report suggested that MeDIP was able to fulfill all these criteria (WEBER *et al.* 2005). However, in attempting to implement this technique according to the published protocol, I came across various problems which led me to try three different protocols: the Weber protocol in the initial publication, an adaptation of this developed in the laboratory of Dr Vardhman Rakyan (VR protocol), and a commercially available kit produced by Diagenode. I will discuss which aspects of the MeDIP technique have proved to be critical, based on my assessments of these various protocols.

Details of the three protocols are in Materials and Methods, but in short there are minor differences between the Weber and VR protocols. The Weber protocol starts with 5µg of sonicated DNA (vs 1.5µg for VR). The beads are washed with different buffers and more beads proportionally are used in the Weber protocol. Following the immunoprecipitation, the DNA is recovered from the beads by proteinase K digestion, the incubation with which is shorter in the VR protocol and at a lower temperature. Following this, DNA is isolated with a kit rather than standard phenol:chloroform extraction in the Weber protocol. The Diagenode kit differs more. The DNA is incubated with both antibody and beads at the same time rather than sequentially (as in the other two protocols). Positive (methylated) and negative (unmethylated) control DNA are "spiked" directly into the reaction, which is carried out with 1µg of starting sonicated DNA. The washes are carried out with buffers the composition of which is unknown. DNA is recovered with a "DNA isolation buffer" to which proteinase K is added for a time shorter than either of the other protocols.

3.4.2 Establishment of a protocol

The MeDIP technique can be broken down into a number of stages: DNA isolation, immunoprecipitation, whole genome amplification, labelling/hybridising/scanning, and data manipulation and interpretation.

3.4.2.1 DNA preparation

Surprisingly, it became evident that an important step is the way in which the starting DNA is isolated. RNAse treatment seems essential to successful immunoprecipitation. Although theoretically the monoclonal antibody is raised specifically against 5-deoxymethylcytosine, anecdotally it has been reported that methylated RNA can interfere with immunoprecipitation of DNA (V. Rakyan, personal communication). This was also an issue here. Figure 3.14 shows a gel of genomic DNA from a number of normal individuals and CLL patients. While MeDIP was carried out on CLL2 and N1, this was only successful in N1. A striking difference between these samples is the presence of large amounts of RNA in CLL2 (as well as most other CLL samples), while this is absent in N1, suggesting this may explain why immunoprecipitation failed in CLL2. When CLL samples were treated with RNAse and MeDIP repeated, this became successful. Thereafter, all DNA samples were RNAse treated.

3.4.2.2 Immunoprecipitation

Clearly, it is critical to have a means of testing whether the immunoprecipitation step has been successful. As there is no such test in the Weber protocol, the approach taken was to devise controls by seeing whether, using RQ-PCR, it was possible to pick up high levels of enrichment in areas expected to be fully methylated, and lack of enrichment in unmethylated areas. Eight RQ-PCR probes were chosen across the 16p locus. Five of these were based in known unmethylated CpG islands (*c16orf35* promoter, *HBA1* promoter and *HBA1* exons 1-2, *HBQ* and *LUC7L* promoter), and three were in the body of a housekeeping gene (*LUC7L*), which would be expected to be highly methylated by virtue of the paradigm of expressed genes having methylated gene bodies (Rauch et al., 2009). Figure 3.15A shows the RQ-PCR result, which did not convincingly demonstrate

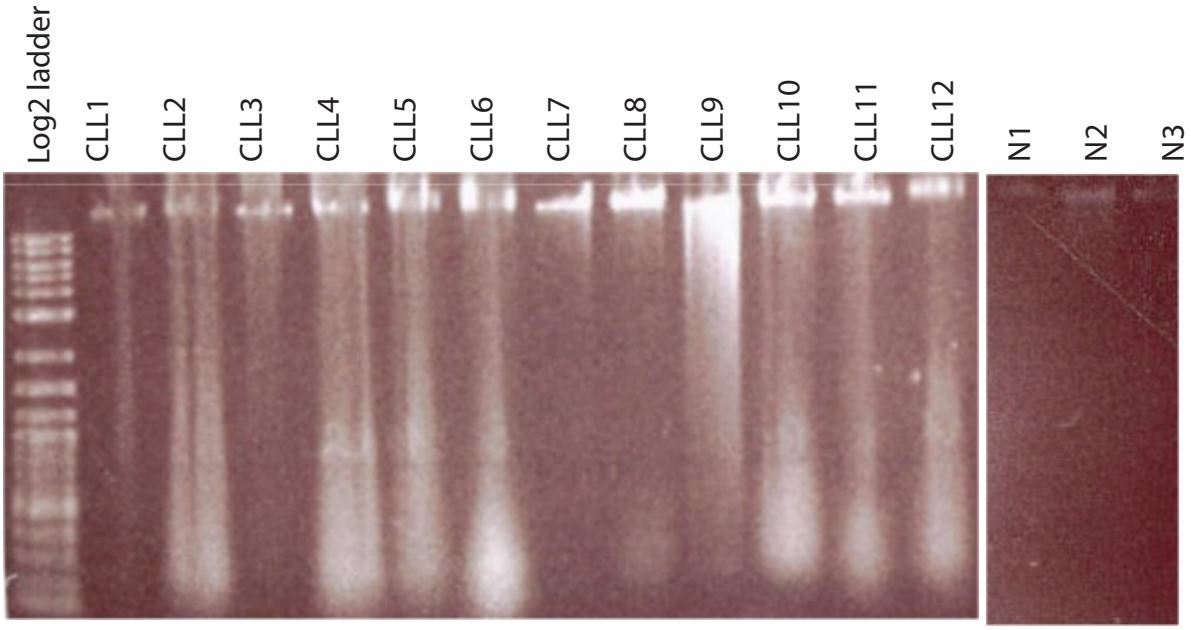


Figure 3.14 Agarose gel of genomic DNA extracted by phenol:chloroform method. Most of the CLL samples show RNA contamination, while the three normals do not.

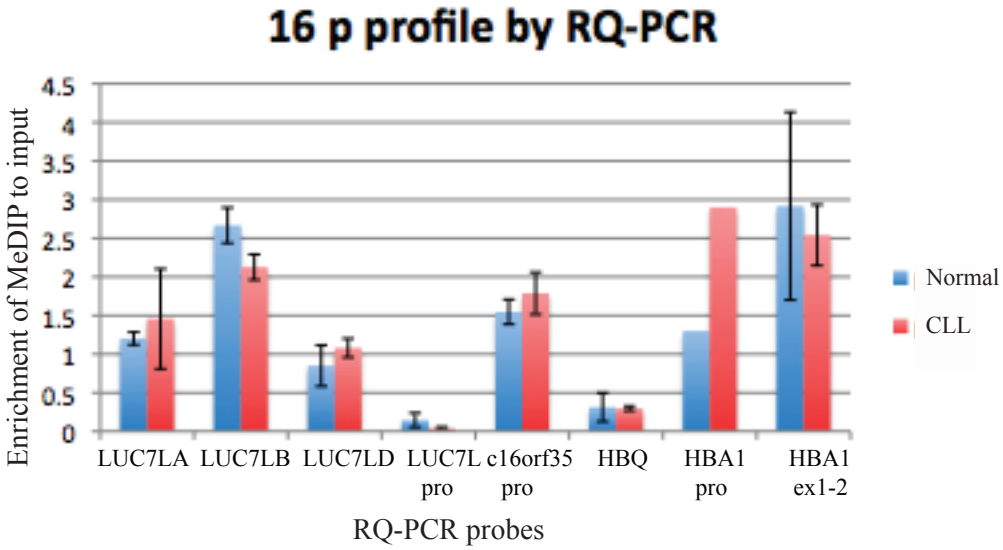


Figure 3.15A RQ-PCR of MeDIP material with probes from 16p locus. Results are expressed as 2 to the power of -(CT MeDIP- CT input). LUC7L A, B and D are within the body of the *LUC7L* gene whereas all the other probes are in CGIs. Enrichment is not low as expected in CGIs, in particular c16orf35 promoter.

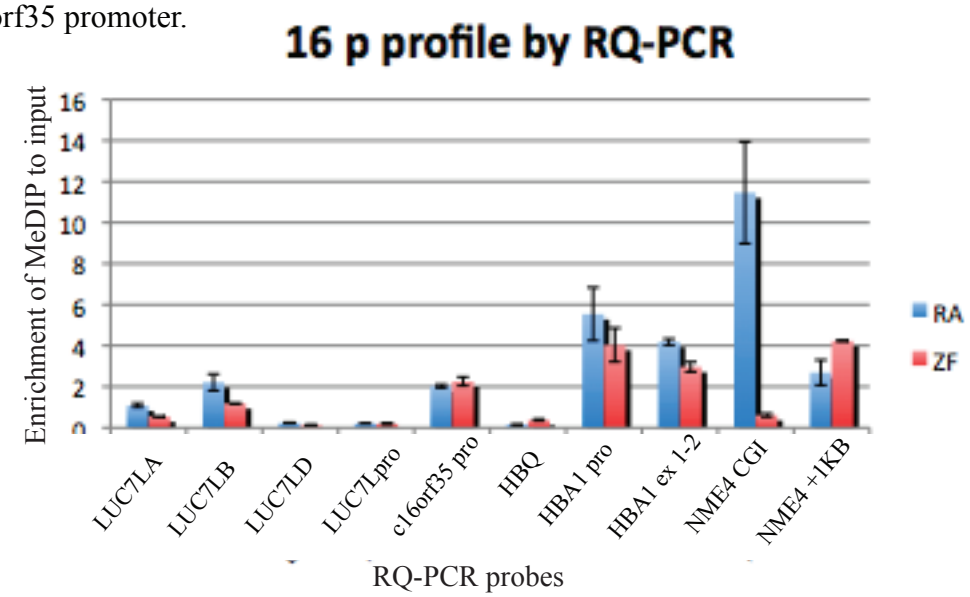


Figure 3.15B RQ-PCR of MeDIP material from RA and ZF with probes from 16p locus. Results are expressed as 2 to the power of -(CT MeDIP- CT input).

Me hylation Status	HEP ID	Chr	NCBI35 coordinates	Description
Me thyla ted	**11821**	20	20574038 - 20574038	LTR repeat
Me thyla ted	**6583**	22	30949610 - 30950048	SINE (MIR)
Me thyla ted	**11851**	20	35510497 - 35510986	AluY repeat
Me thyla ted	**6165**	22	49054726 - 49055151	MER63A DNA repeat
Me thyla ted	**4994**	22	48435751 - 48436242	C22orf34 intron
Unme thyl ated	**9066**	6	13436197 - 13436521	TBCID7 gene
Unme thyl ated	**8804**	6	99886966 - 99887351	ETIFIF4 gene
Unme thyl ated	**5132**	22	30209493 - 30209791	C6orf168

Figure 3.15C MeDIP positive and negative control probes for VR protocol

that the immunoprecipitation step was working reliably, as there was as much enrichment from some of the CGIs as in the bodies of the *LUC7L* gene. The approach then taken was to study two clinical samples, which should provide appropriate controls as they contain regions of abnormally high methylation.

The first control is an alpha-thalassaemia patient (ZF), who has a deletion of *HBA1* and *HBQ1* which results in antisense RNA-induced methylation in the CGI of the remaining *HBA2* gene, abnormally silencing it (TUFARELLI *et al.* 2003). Methylation- sensitive restriction enzyme analysis has shown that this patient has a single fully methylated *HBA* CGI on the affected chromosome, and two unmethylated ones on the unaffected chromosome. This area should therefore show increased enrichment by MeDIP compared to a normal individual.

The second control (RA) is a patient with the inherited ATR-X (X-linked, α thalassaemia mental retardation) syndrome who has been shown to have >90% methylation at the CGI of *NME4*, 220kb downstream of *HBA1* (Richard Gibbons, personal communication). This should also therefore be more enriched by MeDIP compared to normal. The results of RQ-PCR from a MeDIP reaction on RA and ZF with the Weber protocol are seen in Figure 3.15B. Contrary to expectation, the ZF patient does not show higher enrichment at *HBA*. However, RA does have increased enrichment at the *NME4* probe (2-6x that of other amplicons), suggesting that the immunoprecipitation step was indeed successful, at least for that sample. In conclusion, however, one major disadvantage of the Weber protocol is this failure of adequate controls to ensure immunoprecipitation has worked well.

The VR protocol improved upon this. As well as some minor variations with the steps of the immunoprecipitation itself, the main difference is the use of eight control RQ-PCR probes. Five of these correspond to methylated areas of the genome (four repeats and one intron of an unknown gene) and three correspond to unmethylated areas of the genome (three genes), (Figure 3.15C). Once again, however, there was no evidence, even using these controls, that the MeDIP protocol was working reliably. In the example illustrated in Figure 3.16A, two

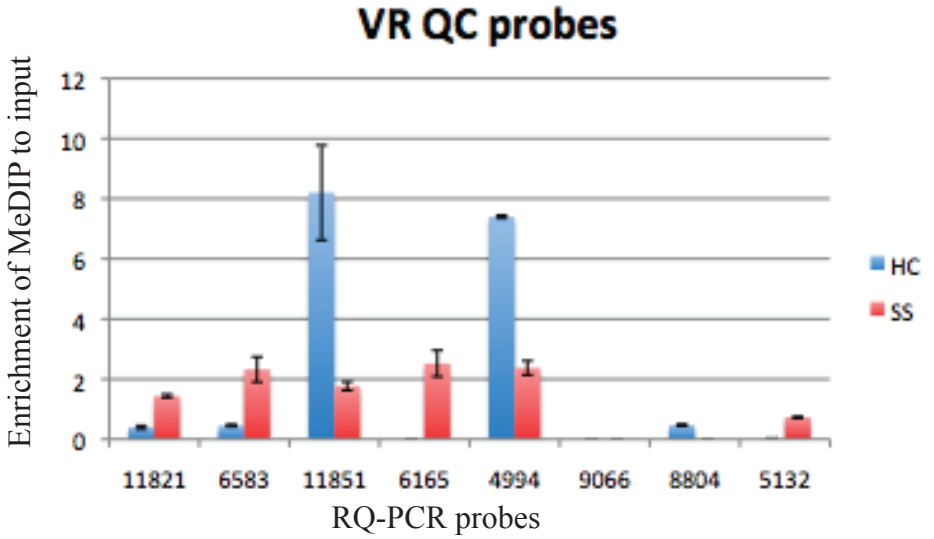


Figure 3.16A RQ-PCR of MeDIP material from two normal samples (HC and SS), analysed with positive and negative control probes. The first five should show enrichment and the last three none.

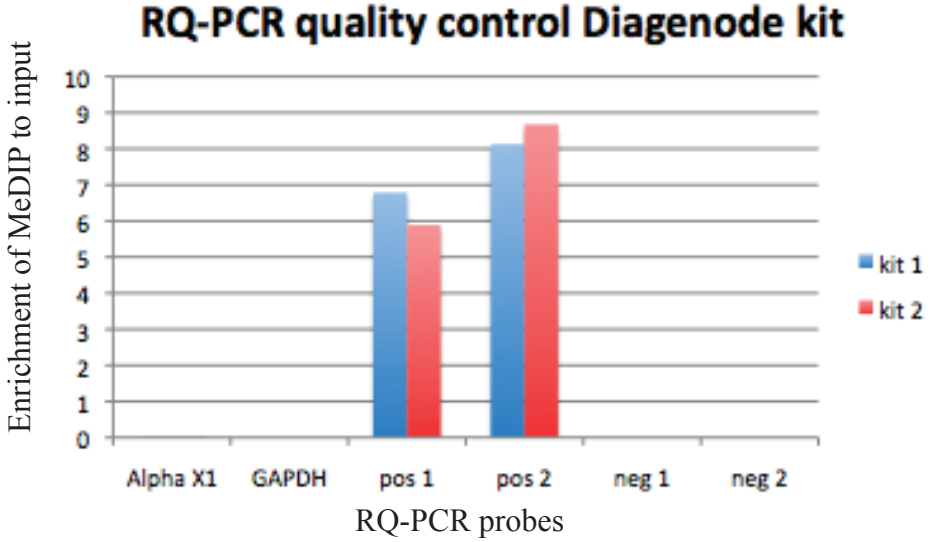


Figure 3.16B RQ-PCR of MeDIP material from a normal sample (duplicated) with Diagenode kit and analysed with quality control probes provided. The immunoprecipitation reaction itself was successful as evidenced by good precipitation of the spiked positive control, but the test DNA has not precipitated as the alphaX1 probe shows no enrichment.

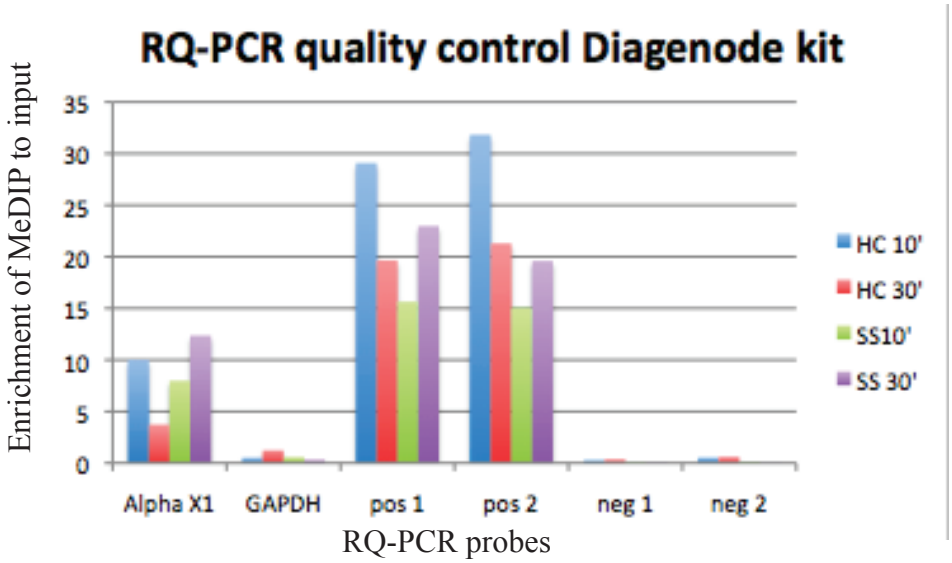


Figure 3.16C RQ-PCR of MeDIP material from two normal samples (HC and SS) under two different sonication conditions. All four samples immunoprecipitated well, as not only did the spiked positive controls enrich, but also HC and SS at AlphaX1.

normal samples (SS and HC) were immunoprecipitated with MeDIP. Sample SS shows enrichment over baseline at all of the methylated probes and none of the unmethylated ones, but the same is not the case for sample HC, which shows enrichment at only two of the five methylated probes. Unless this normal sample happens to be unmethylated at these three repeats, it would seem that these positive and negative control probes are not fully reliable methods of checking whether the immunoprecipitation step has been successful.

This was the main reason for trying the Diagenode kit. Its main attraction is precisely that it provides a reliable method of testing whether the immunoprecipitation step is successful, regardless of whether or not it works on the sample being tested. This is achieved by the use of "spiked in" positive (methylated) and negative (unmethylated) controls. After the MeDIP is completed, it is tested by RQ-PCR using primers for the spiked positive control, the spiked negative control, an alpha satellite repeat (alpha X1) on human chromosome X (which should be methylated) and GAPDH (which should be unmethylated). This allows the discrimination between two scenarios- where the immunoprecipitation failed altogether (no enrichment from spiked positive control or alpha X1) versus the immunoprecipitation process being successful but the sample(s) being tested not working (enrichment from spiked positive control but not from alpha X1, as shown on Figure 3.16B). This has been immensely useful and was a significant reason for selecting the Diagenode kit for further experiments. Figure 3.16C shows positive enrichment from both the spiked controls and the experimental samples.

3.4.2.3 Whole Genome Amplification (WGA)

With the Weber protocol, starting with 5µg of sonicated DNA, about 15ng are recovered following immunoprecipitation. In the Diagenode kit, starting with 1µg, 4ng are recovered. Clearly, this material needs to be amplified before labelling and hybridising, which is carried out by WGA. It was therefore important to assess whether the amplification step was introducing bias, if some areas were being amplified preferentially to others. RQ-PCR was therefore

carried out using the same probes across 16p described above and the CT values for 50ng of sonicated genomic DNA, 50ng of the same material after one round of WGA and 50ng after 2 rounds of WGA were compared. The results are expressed as a proportion of the CT values of the unamplified DNA for each amplicon and shown on Figure 3.17. This shows that the amplified material became detectable at CT values later than unamplified material. This is counterintuitive as there should be more material after amplification than before. However, it reflects the fact that the 16p region being studied is very GC rich and amplifies poorly, so other regions of the genome are being amplified preferentially. The result is that proportionally, there is then less DNA from 16p than from other regions. However, it is reassuring that this is consistent across the region, so although WGA does introduce bias, it appears to do so equally across the entire region under study.

3.4.2.4 Labelling/hybridising/scanning

Once the samples are amplified, the input and MeDIP fractions are differentially labelled with Cy3 and Cy5 respectively, hybridised to a microarray and scanned. The output is a log2 ratio of bound:input signal. However, the scanning step itself was very unreliable. Figure 3.18A shows a .wig file of the same slide scanned at 3 slightly different laser intensities (PMT for Cy5/Cy3: 60/70, 58/68, 62/72), and visualised on the UCSC genome browser. It is worrying that such a small change in laser intensity yields such different signal ratios. When these are plotted on a histogram (Figure 3.18B), it is evident that the output data are vastly different and it is difficult to know which one most clearly reflects the true values. As such, it was decided not to do the labelling, hybridising or scanning in-house but to send the MeDIP samples to Nimblegen for this aspect of the experiments to ensure quality controlled conditions.

3.4.2.5 Data manipulation and interpretation

Once all the data are in the form of log2 ratios of bound:input signal, this still needs to be converted into a quantitative measure of methylation. One

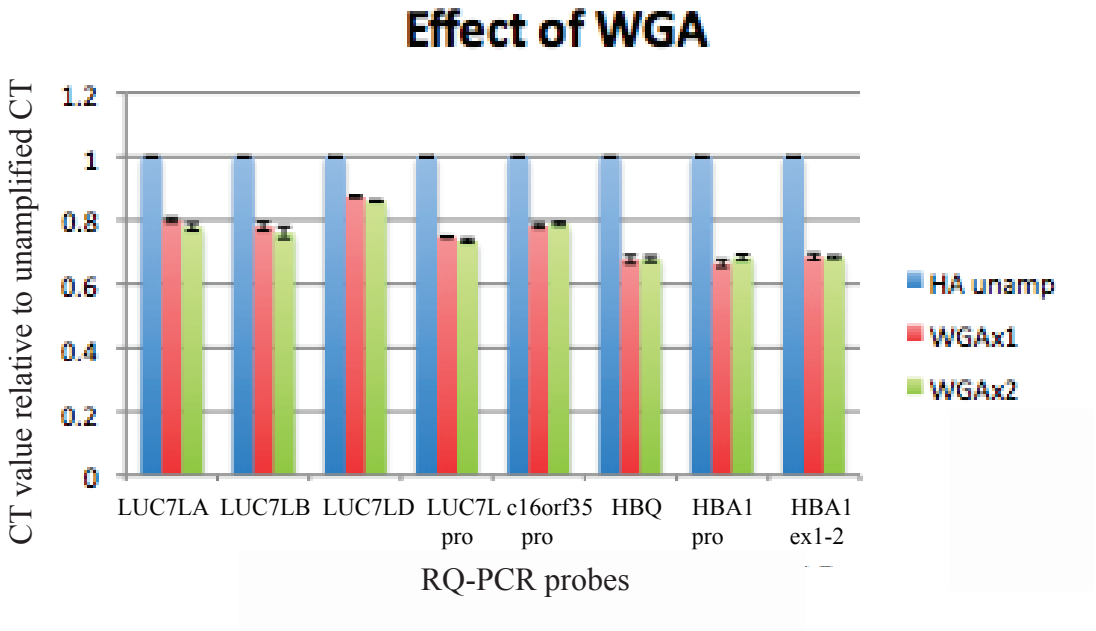


Figure 3.17 Effect of whole genome amplification on amount of DNA amplified by WGA. Sonicated DNA was subjected to one or two rounds of WGA. Fifty nanograms of each WGA material and unamplified ("unamp") material were subjected to RQ-PCR with primers across the 16p locus. The CT values at which each sample crossed the threshold are expressed relative to the CT value at which the unamplified material crosses the threshold for each amplicon. Error bars represent standard deviations from repeating the experiment three times.

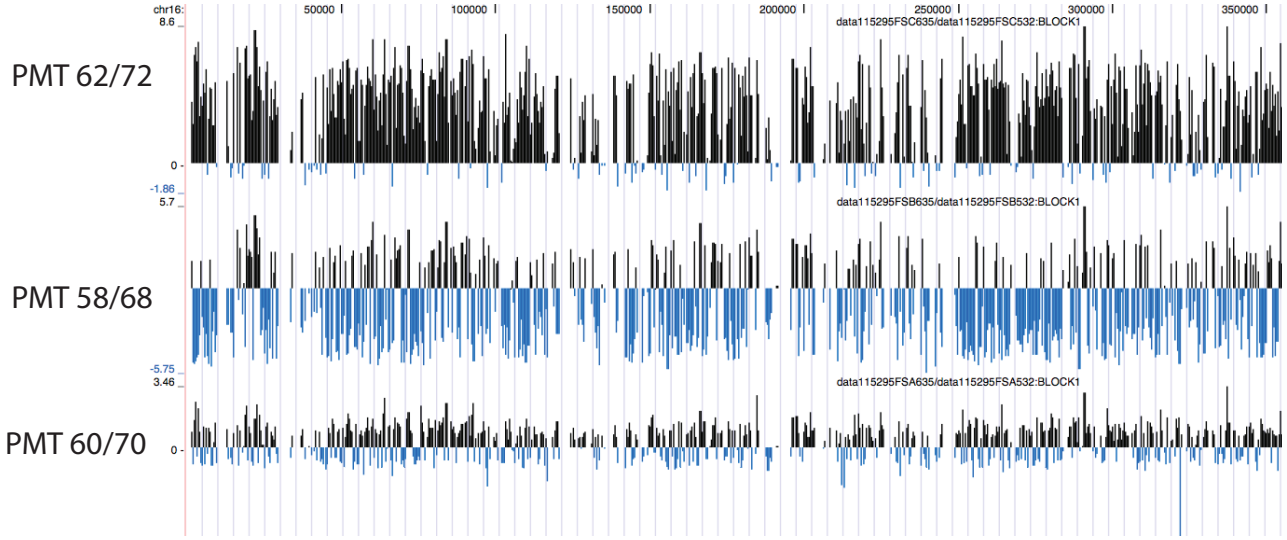


Figure 3.18A Effect of varying laser intensity. Wig files showing signal across chromosome 16:1-500000 and visualised on UCSC genome browser. PMT= photomultiplier tube, which refers to the laser intensity for each of the dyes: Cy5/Cy3.

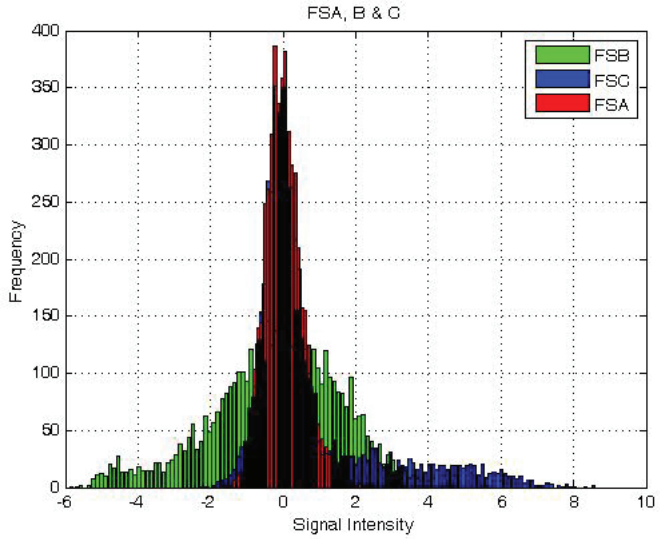


Figure 3.18B Histogram of the frequency of log2 bound:input signals when a single slide is scanned at three different laser intensities. FSA (60/70) gives the most normal distribution. FSB (58/68) is mostly skewed to the left and FSC (62/72) to the right. Without adequate means of normalising for these large differences, nor of knowing what should be the optimal scanning conditions, the data produced will be very dependent on which laser intensity is, arbitrarily, chosen.

important consideration is that the efficiency of the immunoprecipitation will be dependent on the density of methylated CpGs. For example, a fragment containing five methylated CpGs in isolation (i.e. 100% methylated) will immunoprecipitate as well as a fragment containing five methylated CpGs and 20 unmethylated CpGs (i.e. 20% methylation). My initial crude attempts at data interpretation tried to take this into account by dividing the signal ratios produced by the CpG density in each fragment (chosen as the number of CpGs in a 1kb region centered on the fragment in question). When taking this into consideration, the RQ-PCR profile shown on Figure 3.15A is modified to Figure 3.19A. This result is now more convincing, as the enrichment in the gene bodies is now greater than that in any of the CGIs. However, although this calculation was correct in the principle it tried to apply, it was insufficiently sophisticated for the amount of data manipulation required. In order to convert signal intensity to percentage methylation, a complex algorithm (DOWN *et al.* 2008) must be applied to the normalised data. This was developed by Thomas Down at the Gurdon Institute, Cambridge, who kindly agreed to apply his algorithm to my data. This algorithm takes into account the CpG density at each locus on the efficiency of MeDIP enrichment, and converts log2 signal ratios to percentage methylation. A calibration curve is shown on Figure 3.19B. This plots log2 signals on the x axis and CpG density on the y axis. It can be seen that in the low CpG density region (0-10%), the relationship between signal from the array and percentage methylation is linear. However, beyond 10% CpG density, this linear relationship is lost. It is this relationship that is redressed by the algorithm. A MeDIP reaction which has failed (either at the level of the immunoprecipitation itself or at the labelling/hybridising/scanning stage) will not show this linear relationship as seen on Figure 3.19C.

At this point I had settled on a MeDIP protocol (Diagenode kit) which had good reliable internal controls, shown that the amplification step was not introducing significant bias, determined that I would use the Nimblegen service to label and hybridise the arrays, and identified how to analyse the data so that a quantitative measure of methylation could be extracted from the signals produced by the array. I was therefore able to move on to the experimental samples.

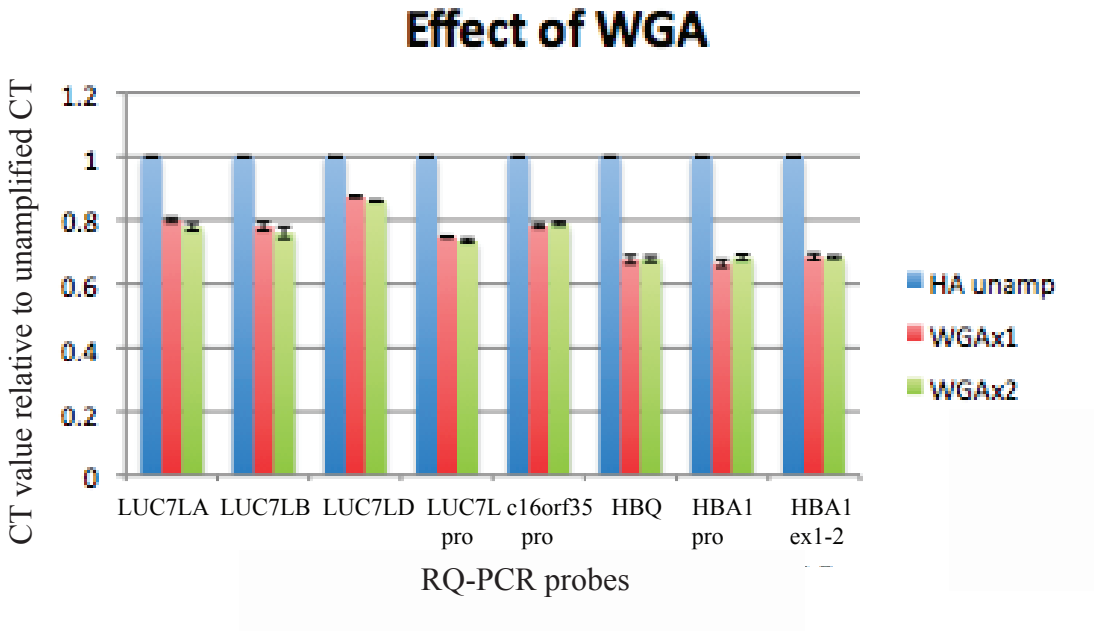


Figure 3.17 Effect of whole genome amplification on amount of DNA amplified by WGA. Sonicated DNA was subjected to one or two rounds of WGA. Fifty nanograms of each WGA material and unamplified ("unamp") material were subjected to RQ-PCR with primers across the 16p locus. The CT values at which each sample crossed the threshold are expressed relative to the CT value at which the unamplified material crosses the threshold for each amplicon. Error bars represent standard deviations from repeating the experiment three times.

3.4.3. Application to an experiment

3.4.3.1 Experimental design

To assess the methylation pattern in CLL, MeDIP was carried out on the whole blood DNA of five normal samples and five CLL samples. This allows pooling of the normals and of the CLLs to reduce the false discovery rate.

3.4.3.2 Quality control

DNA from all ten samples was precipitated with the Diagenode MeDIP kit, following which RQ-PCR was carried out. This showed excellent precipitation of all the samples (Figure 3.20A). These were sent to Nimblegen where they were labelled and hybridised to an ENCODE array (HG18 build).

After Lowess normalization, the Batman algorithm was applied to the data by Thomas Down. The calibration plots showed correct enrichment for all samples except N5 and CLL5 (Figure 3.20B). These were discarded from further analysis. The Batman data were also compared to previously published MeDIP data (RAKYAN *et al.* 2008) from whole blood. Correlations between each of my samples and this reference dataset were 0.4 to 0.6 (see Figure 3.21). This is the same level of correlations as between N1-4 (data not shown).

3.4.3.3 Comparison between MeDIP and bisulfite data

Having both MeDIP and bisulfite data from the same region around chromosome 16 allowed a comparison of the two techniques. For this analysis, all 10 samples subjected to MeDIP were not used, but rather only the two normals and two CLLs for which there was equivalent data by bisulfite sequencing. Despite the fact that the two techniques analyse methylation in very different ways, the correlation between the methylation of individual CpGs as estimated by bisulfite sequencing, and the methylation for the 100bp window corresponding to the same CpG as estimated by MeDIP was assessed. Treating all the CpGs individually, the correlation between MeDIP and bisulfite sequencing is 0.61. As MeDIP assesses methylation over a larger area, it would seem logical that if the CpGs are grouped into discrete regions corresponding to the amplicons

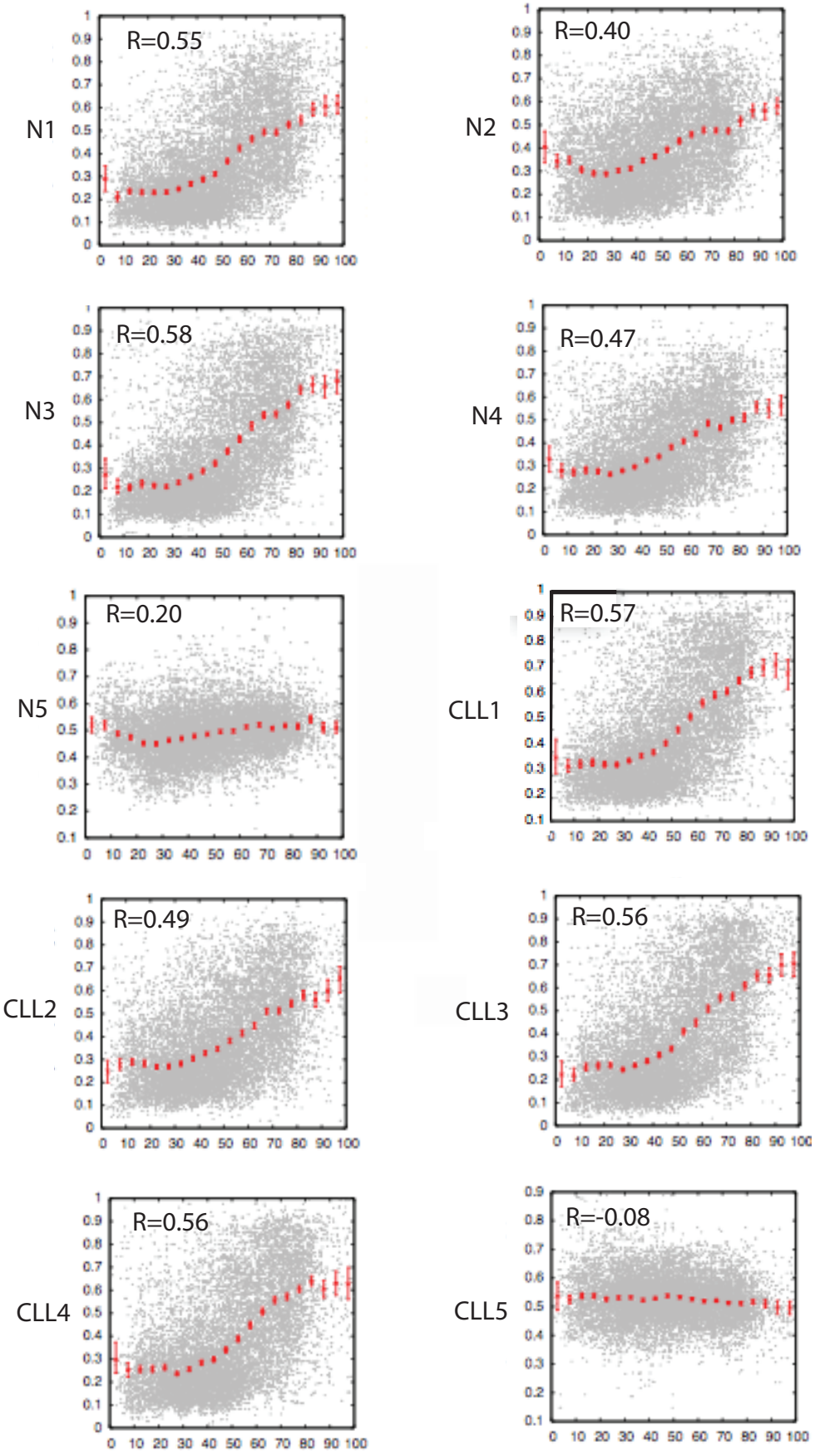


Figure 3.21 Correlations between MeDIP samples (on the y axis) and previously published MeDIP data on whole blood (mPOD, Computational Biology Group, University of Cambridge).

sequenced by bisulfite, and the methylation of all CpGs in one amplicon are averaged, this correlation should be higher. Surprisingly, this was lower at 0.52. The correlation between the two techniques increases if the analysis is done on the average of the two normal samples and of the two CLL samples (0.69 and 0.67 respectively).

3.4.3.4 Differences in methylation between normal and CLL

Although these levels of correlation show a good linear relationship between the two datasets, the critical question for this project is how well MeDIP can pick up differences in methylation between normal and CLL samples. Figure 3.22A shows a track of the average of both the normal and the CLL samples for chromosome 16p by MeDIP, visualised on the in-house 16p browser. An initial analysis of the MeDIP data by Thomas Down looking for areas of extreme difference (<40% methylated in normals and >60% methylated in CLLs and vice versa) identified 11 genes with differential methylation, none of which were in chromosome 16 (Figure 3.22B). However, having identified individual CpGs which become hyper- and hypomethylated in CLL by conventional bisulfite sequencing, this allowed me to compare the ability of these two techniques to pick up abnormal methylation in CLL. Looking at all the CpGs analysed by both bisulfite sequencing and MeDIP, in the same two normal and two CLL cases, a significant difference in the methylation percentage between normals and CLL was sought. For bisulfite sequencing, a greater than 20% difference was used as the cut-off, whereas for MeDIP, a greater than 10% difference was used. This is because of the way that the MeDIP data are distributed; it has a mean of 50% and standard deviation of 14%, such that 95% of the methylation values are between 22%-78% methylation. This lack of values at either extreme and the concentration of values around the mean results in the difficulty of finding large enough differences between any two samples.

Figure 3.23A shows the results of this comparison between bisulfite sequencing and MeDIP for picking up differential methylation in normal vs CLL samples. Twenty-one of the CpGs that are picked up as being significantly different between normal and CLL by bisulfite sequencing are also picked up by the

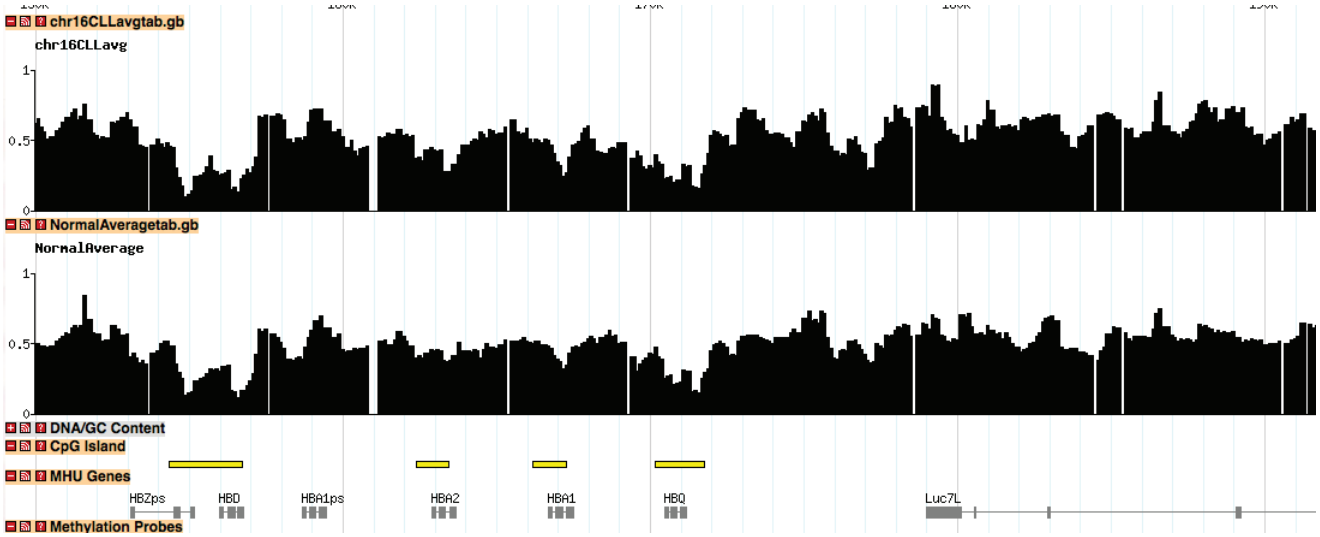


Figure 3.22A Wig files of averages of 4 normals and 4 CLLs for chromosome 16p, visualised on the in-house 16p browser.

Methylation in CLL	Gene abbreviation	Name of gene	Location
Decreased	HBE1	Epsilon globin	11p15.5
Decreased	MDFIC	MyoD family inhibitor domain containing	7q31.1
Decreased	SFI1	Sfi1 homolog, spindle assembly associated (yeast)	22q12.2
Decreased	UGT1A1	Bilirubin UDP-glucuronosyltransferase 1-1	2q37
Increased	BRCC3	BRCA1/BRCA2-containing complex subunit 3	Xq28
Increased	CTSD	Cathepsin D (lysosomal aspartyl protease)	11p15.5
Increased	CTTNBP2	Cortactin binding protein 2	7q31
Increased	HISPPD2A	Histidine acid phosphatase domain-containing protein 2A	15q15.3
Increased	KCRU_HUMAN	Acidic-type mitochondrial creatine kinase	15q15
Increased	MTCP1	Mature T-cell proliferation-1 type A	Xq28
Increased	PFTK1	PFTAIRE protein kinase 1	7q21
Increased	ST7	Suppression of tumorigenicity 7	7q31.1

Figure 3.22B Genes identified as "hypomethylated" and "hypermethylated" in CLL by MeDIP.

	MeDIP <10	MeDIP >10
Bisulfite <20	205	55
Bisulfite >20	42	21

Figure 3.23A Number of CpGs showing differences between normal and CLL samples as picked up by MeDIP and bisulfite sequencing. A greater than 20% difference is used for bisulfite sequencing, but a greater than 10% difference is used for MeDIP (see text for explanation).

Sensitivity	Specificity	PPV	NPV
33.3%	72.3%	22.58%	81.73%
(21-46)	(66-77)	(14-32)	(76-86)

Figure 3.23B Sensitivity and specificity of MeDIP as compared to bisulfite sequencing for picking up differences in methylation between normal and CLL. Values between brackets represent 95 confidence intervals.

MeDIP analysis. However, this leaves out 42 CpGs that are picked up by bisulfite which are not picked up by MeDIP (false negatives for MeDIP). Equally worrying are the 55 CpGs picked up by MeDIP as being significantly different, and which are not so by bisulfite sequencing (false positives for MeDIP). This highlights the limitations of MeDIP. Because the methylation value assigned to any one point reflects the overall methylation of the 500-1000bp around it, it is not sensitive enough to pick up subtle changes that may occur over ~50 bp as can be seen on bisulfite sequencing. Conversely, MeDIP can highlight spurious differences between samples. By reflecting the methylation over larger distances, this 'averaging' can make a region appear to be different between normal and CLL though on bisulfite sequencing of individual CpGs, which provides information at a much higher resolution, no such difference exists. From these comparisons, it is possible to determine the sensitivity and specificity of MeDIP for picking up differences in methylation between normal and CLL samples. This is shown in Figure 3.23B. MeDIP has a sensitivity of 33%, specificity of 72%, positive predictive value (PPV) of 23% and negative predictive value (NPV) of 82%. Clearly, this illustrates that MeDIP is a very poor technique for picking up differences in methylation between normal and CLL samples.

3.4.3.5 CpG islands

Another failing of MeDIP became evident on analysing the methylation at CpG islands, which seemed to be very high even in the normal samples. Islands which were fully unmethylated from bisulfite sequencing, had 30-50% methylation on MeDIP (Figure 3.24A). Further analysis of the methylation levels of 11 CpG islands (mean length 1436bp, range 510-2376bp), showed that, even using the most conservative values for the innermost parts of the islands to minimise spilling over of methylation from adjacent areas, the mean methylation was 28% (range 18-42%, Figure 3.24B). Clearly this is unacceptably high when these islands have been shown to be <5% methylated by bisulfite sequencing. As a result, MeDIP appears to be a poor technique for showing low levels of methylation. As hypomethylated areas are particularly important to identify in the context of malignancy, MeDIP was not suitable for the analyses proposed.

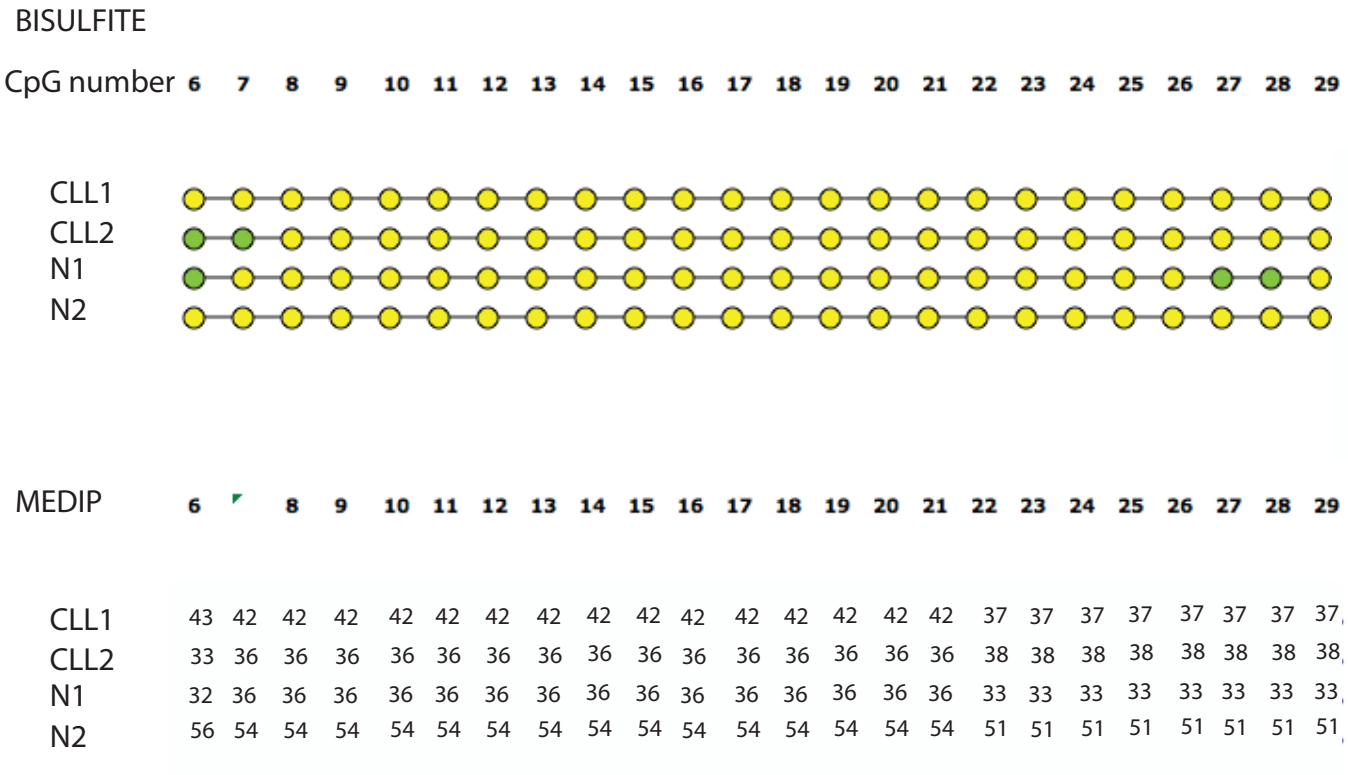


Figure 3.24A Bisulfite sequencing (top panel) and MeDIP (bottom panel- % methylation) data of CpGs 6-29 from the CGI associated with HBA1.

Chr	Coordinates	Mean % meth
16	270062-270995	19
16	314734-315329	27
16	341973-343463	18
16	349986-351492	34
16	162346-162856	42
20	33652134-33653681	32
20	33667093-33668683	27
2	118333046-118334633	34
16	414767-416816	22
16	154344-156720	27
16	371191-372811	21

Figure 3.24B Analysis of the mean methylation in 11 CpG islands by MeDIP

3.5 Discussion

It was initially hoped that using a genome-wide method for studying methylation would allow the identification of differential methylation in CLL in an efficient, reproducible and quantitative manner. However, numerous hurdles had to be overcome in the application of the MeDIP protocol. By attempting to optimise this protocol in different ways, and to find manners of quality controlling the procedure, it became clear which steps of the protocol are crucial to work correctly: the DNA isolation method, the immunoprecipitation itself, and the data manipulation. Having settled on what seemed to be a reliable MeDIP protocol with the Diagenode kit, labelled and hybridised the slides by Nimblegen and carrying out the initial data manipulation to translate the signals into percentage methylation by Thomas Down, the best conditions had been met prior to analysing the results produced.

These were disappointing. When comparing the MeDIP data with that obtained by bisulfite sequencing of the same samples, keeping in mind the different ways in which these techniques work, it became clear that MeDIP has a poor sensitivity and specificity for picking up differences in methylation in CLL compared to normal. This is partly due to the fact that MeDIP assesses methylation over a broader genomic range than bisulfite sequencing, so if the differences in methylation occur at discrete CpGs, MeDIP would not be able to identify these. However, this is not the sole explanation for the poor sensitivity and specificity of the technique. As MeDIP is assessing the methylation over sonicated fragments, which are on average 300bp long, even if it cannot identify low levels of methylation due to discrete hypomethylated CpGs, it should, if it is sensitive, be able to identify low levels of methylation if these occur in a region covering at least 300bp, such as CGIs. However, methylation levels at CGIs as assessed by MeDIP were unacceptably high at nearly 30% for known unmethylated islands.

The concerns highlighted above regarding the sensitivity and specificity of MeDIP have since been echoed by a number of investigators. In one study

comparing MeDIP, HELP and fractionation with McrBC with the GoldenGate Illumina methylation assay (which provides single CpG resolution of selected sites throughout the genome), MeDIP was shown to be an imprecise technique with poor reproducibility of the assay on biological replicates. Furthermore, it found very little difference in overall methylation between an unmethylated and highly methylated cell line. Its correlation with the GoldenGate assay was 0.38 (compared to 0.48 for HELP, 0.63 for McrBC and 0.76 for CHARM-comprehensive high throughput arrays for relative methylation). MeDIP had a low sensitivity for regions of low CpG density (i.e. the majority of the genome) and a worse specificity (IRIZARRY *et al.* 2008a). In another similar study comparing MeDIP-Seq (immunoprecipitation followed by high throughput sequencing), MethylCap-Seq (MBD affinity method followed by high throughput sequencing), and RRBS, MeDIP-Seq was able to identify half as many DMRs as MethylCap-Seq and fewer than RRBS (BOCK *et al.* 2010).

The issues which have been dealt with in the Batman data manipulation protocol were also picked up by another team (PELIZZOLA *et al.* 2008), namely the lack of linear relationship between signal intensity and methylation status, and the efficiency of precipitation varying with CpG density. This group devised a different algorithm for correcting for these variables, based on the immunoprecipitation of a fully (artificially) methylated genome in order to assess the effect of CpG density on precipitation. Using this correction, a correlation of 0.75 between MeDIP and bisulfite sequencing was achieved.

In view of these and the findings presented above, it was decided that MeDIP is too insensitive a method for the purpose of screening for changes in methylation in CLL. The approach was therefore redefined from trying to investigate methylation genome-wide to focusing on methylation changes in the alpha globin cluster. The rationale for using this region as an area to investigate the mechanism behind methylation abnormalities in malignancy is that the alpha globin cluster in CLL is a neutral locus for which there is no selective pressure. The pattern of methylation changes in this region should therefore be representative of changes occurring in other regions of the genome where a

functional effect would then provide the cells with a positive or negative advantage. However, the lack of selective pressure from the alpha globin locus should allow the assessment of methylation changes in an unbiased way. It is hoped that this would provide insight into the deranged mechanisms at play in malignancy and help explain the widespread changes in DNA methylation.

A semi-quantitative bisulfite sequencing assay was used to assess CpG methylation in 330 CpGs across the 16p telomeric region, to allow identification of CpGs whose methylation status changes in CLL compared to normal. Despite the semi-quantitative nature of the assay, the correlations between both two normal and between two CLL samples was excellent, suggesting this is a reliable screening method. Subcloning and sequencing of representative amplicons further confirmed this quantitatively. Nevertheless, while this afforded good quality data from which to draw conclusions regarding the magnitude and direction of changes in DNA methylation in CLL, it remained an attractive aim to obtain highly quantitative methylation at single-CpG resolution, which paved the way for the development of a new technique for targeted capture and high throughput bisulfite sequencing, which will be discussed in Chapter 4.

Chapter 4. High Resolution Assessment of Epigenetic Changes in Chromosome 16p

4.1 Introduction

In order to extend the analysis, a technique was developed to obtain high throughput methylation data at individual CpGs using a targeted capture approach, Targeted Bisulfite Sequencing (TBS). This should theoretically provide highly accurate methylation levels at 2267 CpGs in the 16p region, extending both the accuracy and the breadth of CpGs covered in the previous analyses.

As discussed in Chapter 1, the gold standard method for analyzing DNA methylation would be to generate single-base DNA methylomes by high throughput whole-genome sequencing of bisulfite treated DNA as was first carried out in the *Arabidopsis* genome (COKUS *et al.* 2008). While this was achieved with 110 million reads for *Arabidopsis,* accomplishing the same task for the human genome requires sequencing at a higher order of magnitude. This was first carried out on two human cell lines- hESC and lung fibroblasts, each with 1.2 billion reads (average depth of 14.5 reads) such that 94% of CpGs were covered by >1 read (LISTER *et al.* 2009). The amount of sequencing necessary to produce these results is key to the accuracy and breadth of coverage achieved. For example, in another study, when hESC, an intermediate fibroblast derivative and neonatal fibroblasts were whole genome bisulfite sequenced to study the changes in methylation during differentiation, only 543 million sequences were generated, with an average depth coverage of 3 reads at only 60% of the CpGs (LAURENT *et al.* 2010). The higher resolution achieved by these methods compared to more traditional "genome-wide" methylation techniques has given rise to interesting results on the widespread non-CpG methylation in ES cells (LISTER *et al.* 2009), which has renewed interest in this subject, as well as revealing changes in methylation across splice junctions that suggest a role for methylation in transcript splicing (LAURENT *et al.* 2010).

While the wealth of data generated by these shotgun sequencing experiments allow for an unprecedented detailed assessment of methylation across the genome, the amount of sequencing necessary for adequate sensitivity and specificity renders it costly beyond routine use. Although the cost of high throughput sequencing is consistently coming down, for a depth of coverage comparable to that in the Lister paper, generating about a billion reads per sample is beyond the reach of most laboratories.

Depending on the question to be answered, or merely out of economy, it can be considered adequate to look at DNA methylation with bisulfite modification and high throughput sequencing in a small area of the genome only, so called targeted bisulfite sequencing, bisulfite capture, BS-Seq or MethylC-Seq (LISTER and ECKER 2009). Several versions of this technique have been published, the first of which was applied to the study of nuclear reprogramming. It was based on the technique first developed for exon sequencing, which involves 'padlock' or 'molecular inversion' probes (MIP) containing specific arms that will anneal to the exon, and common arms from which the amplification takes place (MAMANOVA *et al.* 2010). For targeted bisulfite sequencing, MIPs were designed with specific arms to anneal to 2000 CGIs in fibroblasts and iPS cells, once the material had been bisulfite modified (DENG *et al.* 2009). The protocol was not optimal, with 10,000 fold variability in the efficiency of the probes, necessitating dampening down of the signal from the most efficient probes. Another group published a similar method but with MIPs annealing to ENCODE areas and whose hybridising arms selected against CGIs (BALL *et al.* 2009). The process of capture can also be done on arrays (HODGES *et al.* 2009). For all the techniques where capture occurs after bisulfite modification, the design of the capturing probes must take into account the variability of methylation possible at individual CpG sites, which does not necessarily adequately capture partially methylated CGIs if "all/none" methylation is assumed in the design.

The alternative approach, to capture native DNA and carry out bisulfite modification afterwards, requires larger amounts of starting material but could potentially be more accurate for partially methylated CpGs, which is why this

approach was chosen in the protocol described in this Chapter. A region of 76KB in the 16p region comprising the alpha globin locus and surrounding housekeeping genes was targeted for capture of native DNA, followed by bisulfite modification, amplification and high throughput sequencing. Figure 4.1A shows the location of the capture primers in 16p, from positions 94,000 to 172,000. A close-up of a representative area around alpha globin (*HBA2*) in Figure 4.1B shows the spacing between individual probes.

4.2 Targeted Bisulfite Sequencing (TBS)

4.2.2 Establishment of a protocol

4.2.2.1 Overview of the protocol

The Targeted Bisulfite Sequencing (TBS) protocol was designed by Dr Richard Gibbons and initial optimization was carried out by Dr Caroline Scott and Nicola Malik.

Figure 4.2 shows an overview of the capture protocol. A more detailed description of the protocol follows, however, in brief, (1) DNA is sonicated; (2) It is then denatured and hybridised in solution with the library of biotinylated capture oligos. The captured material is recovered with streptavidin beads; (3) The efficiency of capture is assessed by RQ-PCR using primers within and outside the capture region; (4) If the capture has been successful, the material is then bisulfite converted; (5) This is then amplified using random primer amplification (RPA) with primers containing an AcuI digestion site; (6) These primers are then trimmed off with AcuI digestion; (7) The material is size selected and gel purified; (8) A small amount of recovered DNA is subcloned and sequenced to assess whether the material recovered is bisulfite converted; (9) If good quality bisulfite converted material is obtained, the remainder is processed by high throughout sequencing on an Illumina platform.

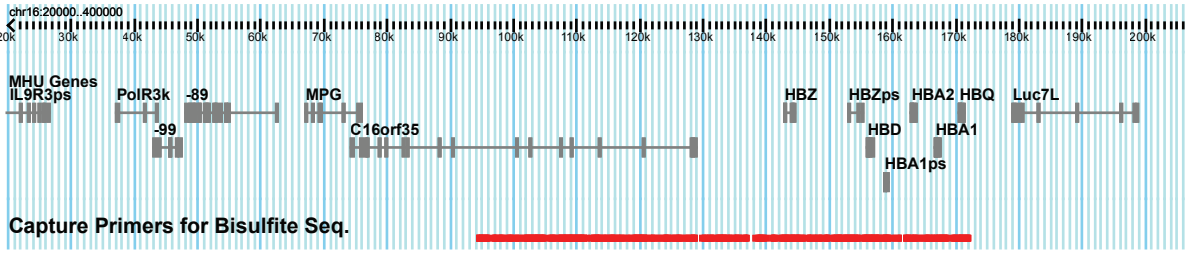


Figure 4.1A Snapshot of the capture area, showing which genes are included. The capture oligonucleotides are tiled at regular intervals in the 76KB region. Red bars indicate the position of the oligos.

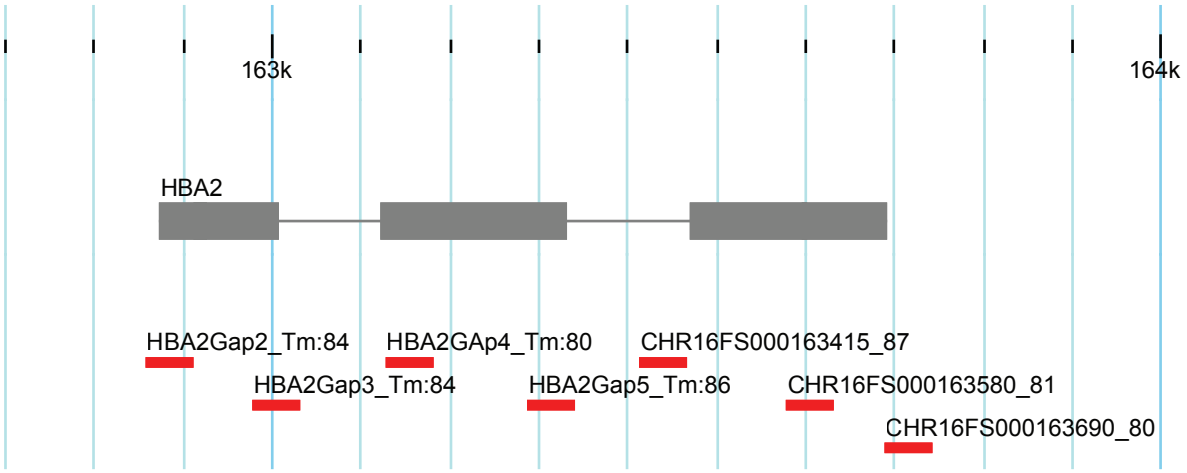


Figure 4.1B Zoom into 1kb of the capture region, showing alpha globin (*HBA2*) and illustrating the distance between the capture oligos. Red bars indicate the position of the oligos.

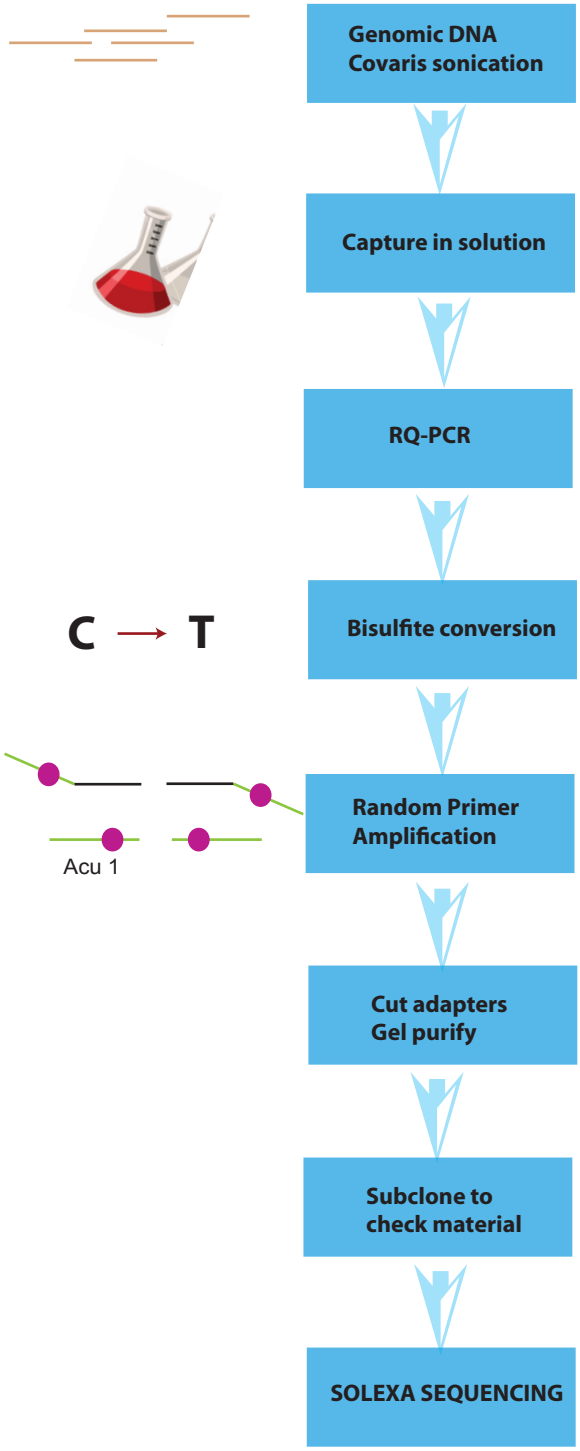


Figure 4.2. Overview of the capture protocol. DNA is sonicated to an average length of 500bp, denatured and hybridised in solution with the library of biotinylated capture oligos and the captured material recovered with streptavidin beads. RQ-PCR is used to assess the efficiency of the capture, after which the material is bisulfite converted and amplified using random primer amplification (RPA) with primers containing an AcuI digestion site. The primers are then trimmed off with AcuI digestion, the material size selected and gel purified. A small amount is subcloned and sequenced and if good quality bisulfite converted material is obtained, the remainder is processed by high throughout sequencing on an Illumina platform.

4.2.2.2 Optimisation of individual steps

(i) Sonication

The initial starting material is 5µg of genomic DNA. This is sonicated to fragments with an average length of 500bp using the Covaris Adaptive Focus Acoustics. The controlled wavelength produces more uniform shearing without the "hotspots" that result from traditional sonication. The DNA is run out on an agarose gel to confirm adequate sonication.

(ii) Capture

This is carried out in solution with a 100x excess of oligo library to capture area. The library comprises 456 oligos with an average length of 52bp. The oligos were designed by Richard Gibbons and Jim Hughes, with the majority being based on the sequence from Nimblegen's high density tiled arrays, which are designed to be isothermically optimised. The remainder of the oligos were manually designed to match the length and melting temperature of the exisiting oligos. They cover an area of 23.7kb distributed across 76kb. The sequence corresponds to the top strand, such that only the bottom strand should be captured. Repetitive regions are excluded. The oligos are biotinylated at the 5' end. Following denaturation, the capture occurs at 65°C overnight in the presence of hybridisation buffer. The next day, the captured material is recovered using streptavidin-coated Dynabeads and washed with an SSC/SDS buffer. The material is eluted with NaOH, then desalted with a G50 column and resuspended in water.

(iii) RQ-PCR

To ensure that the capture has been efficient, an RQ-PCR is carried out with eluted DNA. Primers inside and outside the capture region are used for both captured and input DNA. Enrichment levels are calculated as $2^{\Delta\Delta CT}$ (capture:input for the primers inside the capture area vs the primers outside the capture area). Adequate enrichment was deemed to be in the order of 5000-10000 fold. Figure 4.3A shows a typical RQ-PCR enrichment plot for two samples, 10A and 10B.

Figure 4.3A Enrichment levels following capture. Fold enrichment is calculated from RQ-PCR as $2^{\Delta\Delta CT}$ (capture:input for the primers inside the capture area vs capture:input for the primers outside the capture area).

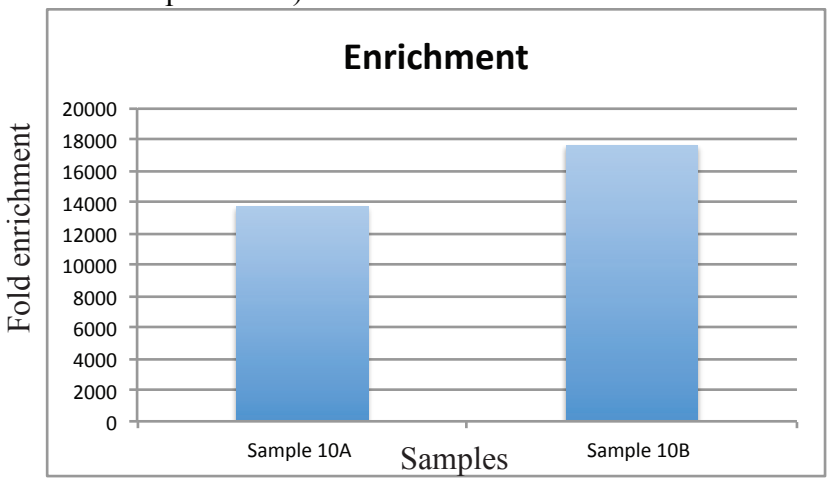


Figure 4.3B AcuI digestion. The digested primers are easily identified at the bottom of this 1% agarose gel.

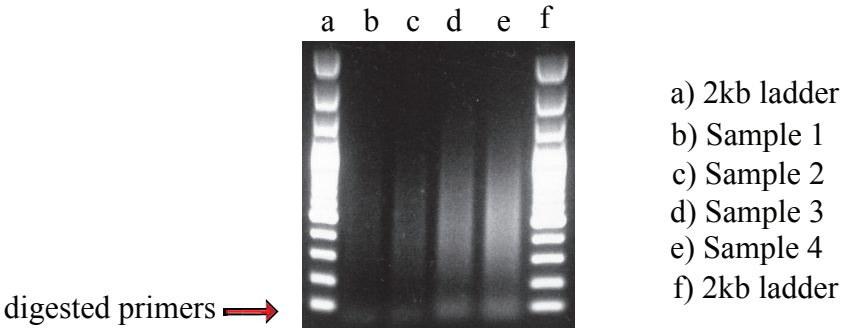
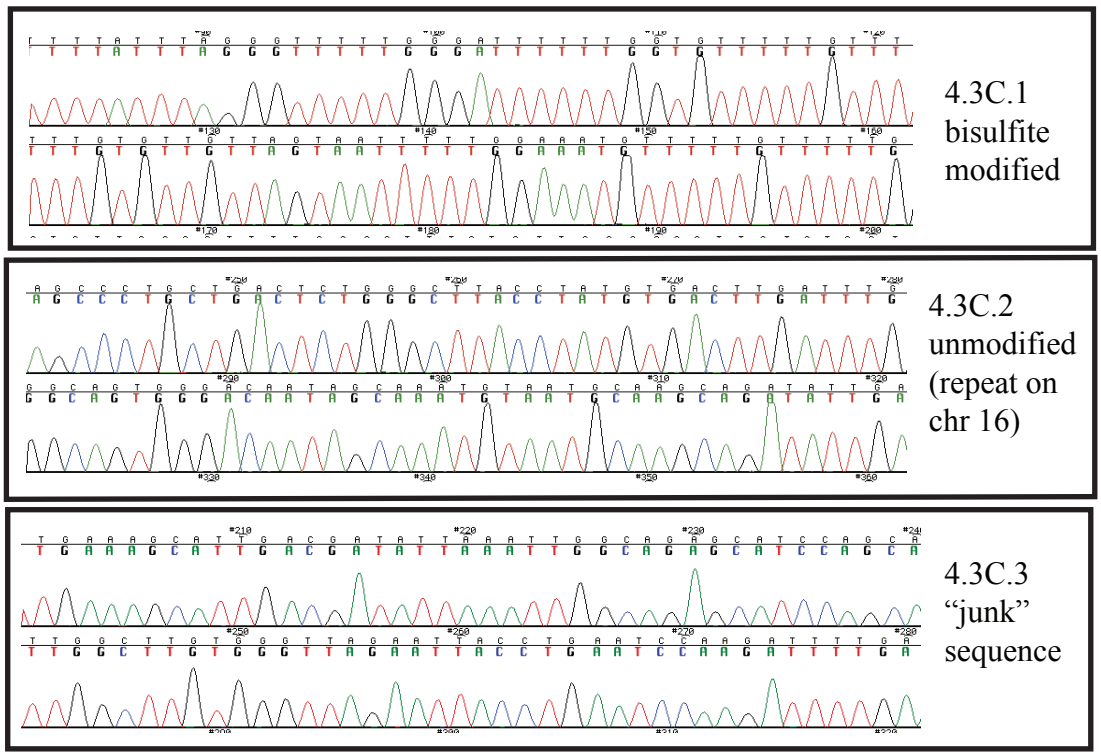


Figure 4.3C Sanger sequencing of subcloned material



(iv) Bisulfite conversion

Initial optimization carried out by C. Scott compared three different bisulfite conversion kits (MethylXL, EpiTect, Zymo). All three are reported to have conversion rates of >99%. There can be two types of non-conversion: where the entire strand has failed to convert (e.g. fragment not denatured), and where individual Cs have failed to convert in a strand where the majority of Cs have successfully converted (e.g. local conformation obscures part of the sequence). In traditional bisulfite sequencing, where bisulfite conversion is followed by nested PCR with primers specific for bisulfite converted DNA, the first type of non-conversion is of no consequence as unconverted strands will not amplify. The second type of non-conversion is more problematic as an unconverted C in a CpG context will erroneously return a methylated CpG in the analysis. However, in the capture protocol, entire unconverted strands would also be amplified as the first round of amplification occurs with degenerate primers (see below). These would be sequenced and introduce bias in the methylation calls. Efficiency of conversion was therefore not calculated as number of unconverted Cs, but number of unconverted strands as assessed by subcloning and sequencing. MethylXL gave the best conversion rates and was selected for the protocol (e.g. 50-100% for MethylXL vs 25-67% for EpiTect, C. Scott).

(v) RPA (Random primer amplification)

The yield from the capture process is extremely low. Assuming 100% efficiency, from 5µg of genomic DNA, capturing 76kb from 3,000,000kb in the entire genome would yield 0.12ng DNA. As bisulfite conversion further fragments and causes loss of DNA due to the high temperatures and low pH necessary, following this step there is not enough material from which to construct a library for sequencing. It is therefore necessary to amplify the DNA. This was done with a method based on random primer amplification. Two rounds of PCR are carried out with primers that carry an AcuI digestion site for later cleavage of the primers, as well as a degenerate region (RG3A, see Appendix 1.E). The number of degenerate bases was reduced to 6 in accordance with the literature that has shown a reduction in amplification bias if fewer degenerate bases are used

(REINDERS *et al.* 2008). Due to the degenerate region, these will bind unmodified DNA as well as modified, which will introduce unconverted DNA into further steps. Following the initial rounds with the degenerate primers, further rounds were carried out with primers that anneal to the specific arms of RG3A (RG4, see Appendix 1.E). PCR bias is minimised by limiting the number of cycles and carrying out the PCR in several tubes and pooling the results. Subcloning and sequencing a fraction of the material controls for quality. Initially, a high proportion of "junk" sequences were noted, which were neither bisulfite converted sequences (easily recognizable as there are no Cs, Figure 4.3C.1), nor unconverted sequences that map to the genome (Figure 4.3C.2, in this example a sequence mapping to a repetitive element in chromosome 16), but sequences that look unconverted (contain Cs) but do not map back to the genome (Figure 4.3C.3). These likely correspond to DNA generated from self-priming of the degenerate oligos. To reduce these to a minimum, the amount of RG3A primers was reduced from 40μM to 4μM.

A combination of polymerases was chosen for the RPA reactions to exploit their different properties. Sequenase has high processivity and excellent displacement properties, which reduces the amplification of small fragments that result from annealing of the non-specific oligos to the middle of strands. AmpliTaq Gold is thermostable and PWO has extremely low error rate and excellent proofreading capacity.

(vi) AcuI digestion

The use of the RPA primers introduces a further 17 bases to each captured fragment. High throughput sequencing with 51nt would therefore waste a third of the sequence, in particular the beginning, which is usually of the highest quality. In order to avoid this, an AcuI digestion site was included in the primer sequence (see Appendix 1.E). AcuI is a type II restriction enzyme that removes 22 bases from its recognition site, ensuring that no primer will be sequenced. The digested primers are easily identified on an agarose gel (Figure 4.3B).

(vii) Gel purification

Following digestion, the material was extracted from low melting point TAE agarose gels. As discussed above, there was a high frequency of "junk DNA" from the initial protocol. Although this was reduced by altering the RPA conditions, they could not be altogether removed due to the necessity of using degenerate primers. However, these non-specific sequences are more likely to be small in length (<200bp), therefore to reduce their sequencing to a minimum, only material >300bp in length was excised from the gel and then purified using a standard kit (Qiagen).

(viii) Subcloning and Sanger sequencing

As just discussed, this essential step is carried out to verify the quality of the DNA prior to sending it for sequencing and has allowed optimization of the protocol and selection of the best material for high throughput sequencing.

4.2.3 High Throughput sequencing

High throughput sequencing

This was carried out at the Wellcome Trust Centre for Human Genetics (WTCHG) according to the standard protocol as described in Chapter 2.

4.2.4 Data manipulation

Following high throughput sequencing, the data must be processed through an elaborate pipeline before methylation information can be determined. The process is illustrated in Figure 4.4 and follows the sequence of: (1) Fast QC; (2) Bismark; (3) Methylation Extractor; (4) Seqmonk. These steps will be described in turn.

Processing of the data received from sequencing until input into Seqmonk was done by Nicki Gray (MRC Molecular Haematology Unit, Weatherall Institute for Molecular Medicine).

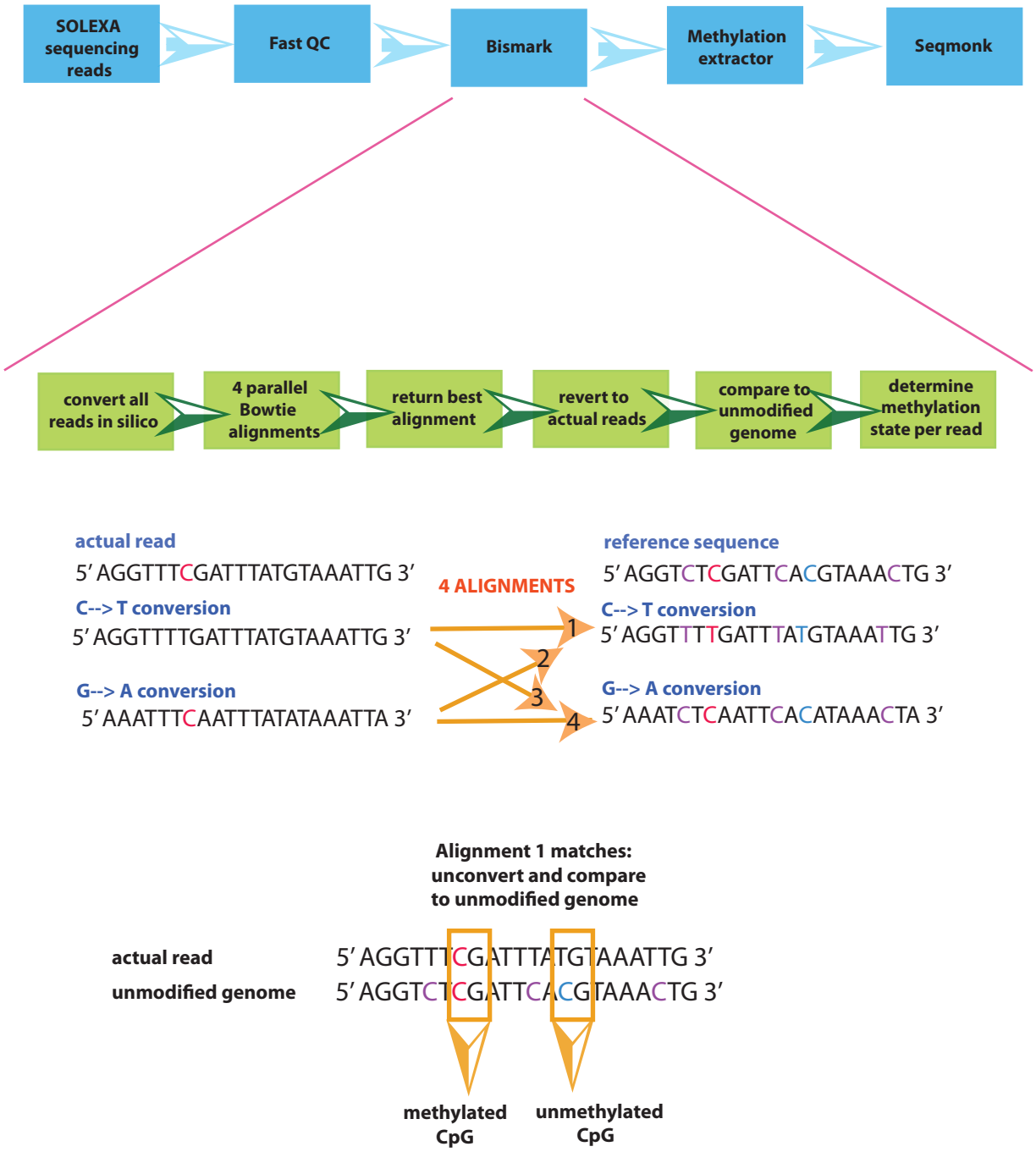


Figure 4.4 Data analysis pipeline. The sequencing reads are quality controlled (Fast QC) and then taken through to Bismark for mapping. This process involves converting all reads *in silico* with both a C to T conversion and then a G to A conversion, aligning each of these to the sequences of the equally converted reference genome, recording the position of the mapped reads, reverting to the actual reads and comparing them to the unmodified genome. Where CG matches, a methylated CpG is recorded and where TG matches, an unmethylated CpG is recorded. The methylation levels of all CpGs for all reads are taken through to the methylation extractor, which, based on the number of CGs vs TGs in all the reads for individual CpGs works out a methylation % at each CpG for which sequencing data is available. This data is then exported into Seqmonk where methylation levels can be visualised and the data further manipulated, for example selecting only CpGs for which a minimum number of reads is available.

4.2.4.1 Fast QC

This is a platform for carrying out quality control on the reads generated by the Illumina sequencer. It ensures that the raw data are of sufficient quality for carrying on with analysis and that there are no biases introduced by the library preparation or the sequencing itself. There are several aspects to the Fast QC process, not all of which are applicable to bisulfite modified material (for example, GC content and base composition are not appropriate). One of the most useful features is the "per base sequence quality", which scores the quality across all bases at each nucleotide position. Typically, quality decreases towards the end of the reads. It is possible to identify whether a subset of the sequences all have a low quality (such as being poorly imaged due to their position in the flow cell). Another feature of Fast QC is the identification of over-represented sequences, which could come from the adapters used to prepare the library, or from contamination with foreign DNA. An over-represented sequence is defined as one which makes up more than 0.1% of the total.

Once the sequences are confirmed to be of adequate quality, they can be taken through for mapping.

4.2.4.2 Bismark

Mapping bisulfite converted sequences to the genome poses additional challenges to the usual task of mapping short high throughput sequencing reads. Due to the vast depletion of Cs from the reads, the complexity of the sequences is greatly reduced, which compromises unique alignment. Secondly, the search space is significantly increased, as the forward and reverse strands are no longer complementary following bisulfite modification. Thirdly, a T in a read could map back to the original genome as a T or a C but not vice versa (asymmetric mapping), which further complicates alignment. Furthermore, in algorithms that mask the C/T bases, the computational load is greatly increased, as all mapping positions must be recorded, even non-unique ones (whether they are unique and should be kept or non-unique and should be discarded will only be revealed after the masking is reversed, which is after the mapping process). Finally, the possible heterogeneous methylation of individual CpGs (which would be

reflected as CG vs TG in reads) necessitates a degree of mismatching to the original genome.

The ideal mapping strategy would be to align the reads while allowing each T to map to *either* C or T. This was the strategy employed for the *Arabidopsis* genome (COKUS *et al.* 2008), but this is extremely computationally intensive and currently impossible to carry out for larger genomes. The choice of algorithm will depend on available computational time and memory, and need for accuracy. In general, increasing read length and doing paired end mapping will greatly increase accuracy. Initial bisulfite mapping algorithms such as BISMA (ROHDE *et al.* 2010), BSMAP (XI and LI 2009) , SOAP2 (LI *et al.* 2009), and BRAT (HARRIS *et al.* 2010) were not only relatively inaccurate and required a lot of computer memory, but once the mapping had been carried out, further data processing was necessary for determination of methylation levels. Newer algorithms such as BS-Seeker (CHEN *et al.* 2010) and Bismark (KRUEGER and ANDREWS 2011) outperform these in memory, time and accuracy, and incorporate determination of methylation levels within the software. Bismark is the least likely to return non-unique reads, at the cost of a small increase in computational time.

Bismark was therefore chosen for the downstream analysis. This is an algorithm that has been specifically designed at the Babraham Institute by Felix Krueger for mapping bisulfite converted reads. It is based on Bowtie, an ultra-fast memory efficient short read aligner (LANGMEAD *et al.* 2009). It is illustrated in Figure 4.4 (bottom panel). The steps are (1) Convert all reads *in silico*, (2) run 4 parallel Bowtie alignments, (3) return best alignment, (4) unconvert all reads (i.e. revert to the original reads), (5) compare to reference genome, (6) methylation extractor, all of which are detailed below.

4.2.4.2.1 Convert all reads in silico

Figure 4.5 illustrates how, after bisulfite conversion and PCR amplification, each double-stranded sequence of DNA will give rise to one of 4 different strands:

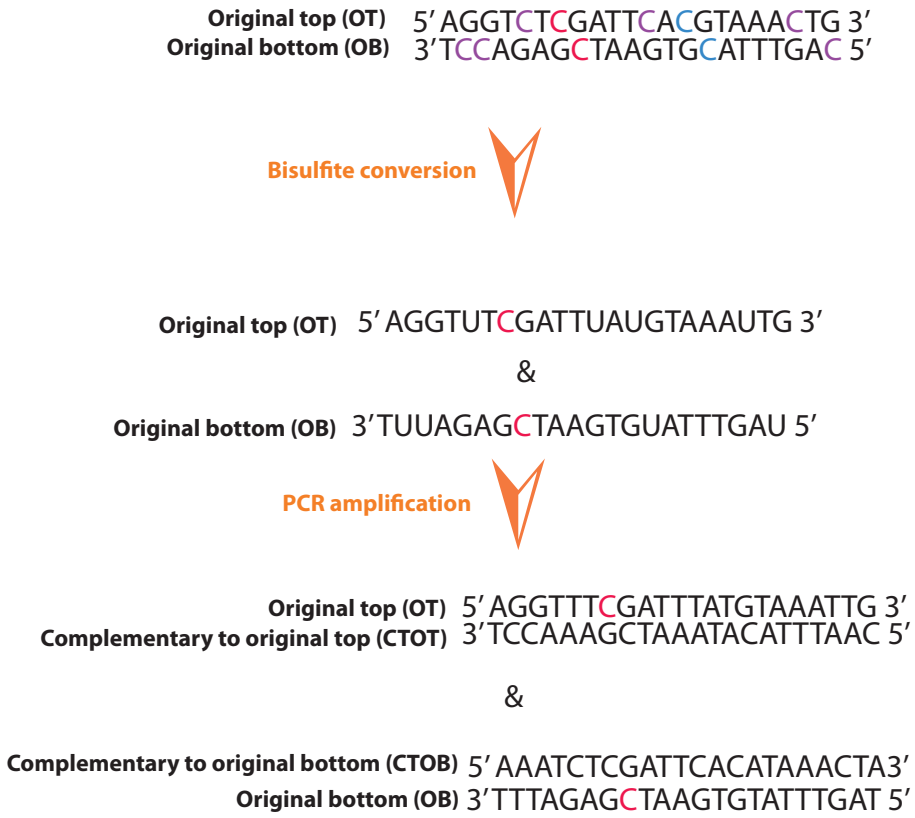


Figure 4.5 Example bisulfite modification. From a double-stranded segment of DNA containing an original top (OT) and an original bottom (OB), bisulfite conversion creates two non-complementary strands, which after PCR amplification yield OT and CTOT (complementary to original top) and OB and CTOB (complementary to original bottom), a total of 4 distinct strands. .

original top (OT), complementary to original top (CTOT), original bottom (OB) and complementary to original bottom (CTOB). Therefore a read from the high throughput sequencing could come from any of those 4 strands, although as explained above the probes were designed to capture only the bottom strand, so it would be expected that the reads should come from the bottom strand or its complement. The read is converted in two separate ways: all Cs to Ts or all Gs to As.

4.2.4.2.2 Run 4 parallel Bowtie alignments

Each converted read is then allowed to try to align to one version of the genome where all Cs are converted to Ts and one where all Gs are converted to As. Therefore for each read, there will be 4 separate mapping processes as illustrated in Figure 4.4 (bottom panel).

4.2.4.2.3 Return best alignment

The best alignment from the 4 mapping processes is returned for each read, and by definition the strand is specified. Sequences aligning at multiple sites are discarded.

4.2.4.2.4 Unconvert all reads

Following mapping, the sequence of each read is reverted back to the actual sequence from Illumina and the mapping position of each read recorded.

4.2.4.2.5 Compare to reference genome

The unconverted read (the sequence determined by the Illumina sequencer) is compared to the reference genome. Where a CG is returned, this corresponds to a methylated CpG and where a TG is returned, this corresponds to an unmethylated CpG.

4.2.4.3 Methylation Extractor

Every read now has every CpG scored as methylated or unmethylated. The methylation extractor aspect of Bismark records the methylation state of individual CpGs by taking into account all reads that correspond to a single CpG, and enters their position and methylation state for uploading into Seqmonk. All

sequence information is discarded and only CpG position and methylation state retained.

4.2.4.4 Seqmonk

The information that is uploaded to Seqmonk therefore records CpG positions and its state as methylated or unmethylated. In Seqmonk it is possible to visualise the data as well as carry out quantitation such as percentage methylation, and to filter on position or read count. The analysis which follows was carried out on CpGs for which there are at least 10 reads.

4.2.5 Initial results: technical issues

Once the protocol had been devised and each step optimised as best possible from assessment of the material that was subcloned and sequenced, initial samples were processed for high throughput sequencing. The first two samples to be sequenced were initially used for further assessment of the technique and optimization of the protocol and data analysis pipeline.

4.2.5.1 Fast QC results

Two samples, N1 and CLL1, were initially prepared and sequenced in two independent flow cells. N1 was sequenced first and the Fast QC showed that the quality of the bases from 51-100nt was poor (see Figure S4.1A in Appendix 2) and the sequences were therefore trimmed to 50bp for further analysis. The poor quality of the reads was attributed to the reduced complexity that results from sequencing a bisulfite modified genome. This means that the software that carries out quality control does not see the equal spread of bases that it expects and therefore is unable to create the matrix that allows quality scoring. All subsequent sequencing was therefore carried out to 50bp only. In the case of N1, there was also an over-represented sequence (Figure S4.1B), but its origin could not be determined as it did not match any primers or adapters used, nor did it map to any known genome. This sequence made up 0.12% of the forward reads and 0.24% of the reverse reads. For CLL1, while the quality of the forward sequences was good, that of the reverse sequences was so poor as to be unusable (Figure S4.1C). There was also an excess of overrepresented sequences (63 in

total). A number of these is shown on Figure S4.1D. In total, they make up 22% of all the reads sequenced. They correspond to the Illumina adapters used to make the library. Even for the sequences for which "no hit" is reported, the sequence is clearly derived from the adapters and represent sequencing errors. There is a direct link between the over-represented sequences in the forward reactions and the poor quality in the reverse reactions, which again relates to the reduced complexity that comes from sequencing the same sequence repeatedly (the adapters), which compromises adequate quality calls on the subsequent run. As the reverse reads could not be used for analysis, the data had to be analysed as single end rather than paired end. This had implications for further samples sequenced, even those which had good quality forward and reverse reads, as to make all the data comparable, the analysis had to be done as single end for all the samples sequenced. Possible reasons for the excess of adapters in the sequencing include too little DNA or poor ligation due to an inhibitory reagent in the final elution. It was therefore decided that the Wellcome Trust Centre for Human Genetics (WTCHG) would re-measure the DNA and carry out an extra cleanup step prior to doing the library preparation of the next samples, which was done with Ampure beads (Agencourt).

Initially, the reads were mapped with Bismark allowing for three mismatches (3mm). This was decided in order to allow the maximum matching of reads that contain several CpGs which could have any combination of methylation/unmethylation. However, a detailed look at the sequences mapped revealed several concerns, which will be discussed in turn:

(a) Poor mapping efficiency
(b) Inaccurate mapping
(c) PCR bias
(d) Strand discrepancies
(e) CpG density bias

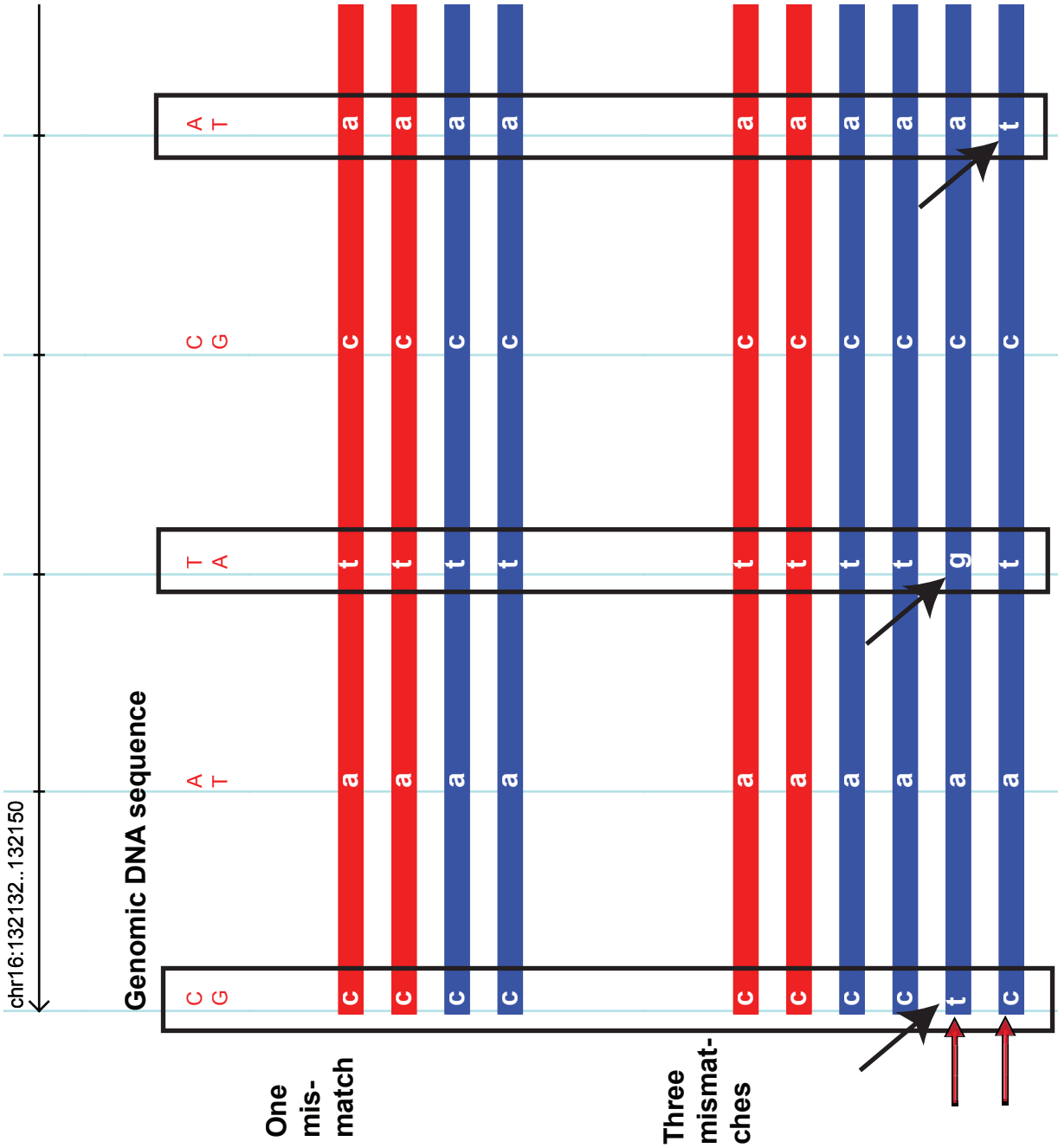
4.2.5.2 Poor mapping efficiency

For sample N1, mapping efficiency was 5.5% allowing for three mismatches. Out of 2267 CpGs in the capture region, 1200 were informative at more than 10 reads. Surprisingly, only about 5% of reads mapped, compared to an estimated 55-65% from previous reported use with the Bismark algorithm (KRUEGER and ANDREWS 2011). This is probably due to the 'junk DNA' created by the RPA (Random Primer Amplification). An attempt was made to improve on this by reducing the number of PCR cycles in RPA-B from 32 to 30. This had the effect of reducing the amount of 'junk DNA' on subcloning from 36% and 50% for N1 and CLL1 respectively to 18% and 0% for N2 and CLL2 (see later), and a concurrent improvement in mapping efficiency to 14.2-16%.

4.2.5.3 Inaccurate mapping

Allowing three mismatches was too low a stringency as it allowed inaccurate mapping of a number of reads. In Figure 4.6, the bottom panel shows the reads mapped when allowing three mismatches. Two of these reads (marked by red arrows) are misaligned, as shown by the mismatched bases (highlighted by the black arrows and black boxes). This degree of mismatch would allow too many incorrect sequences to contribute to the methylation calculations and therefore the stringency was increased by allowing only one mismatch per read. While this had the effect of removing inaccurate reads (the misaligned reads are no longer seen when the algorithm is run again with only one mismatch allowed, as shown on the top panel of Figure 4.6), it also caused a significant loss of reads. For example, the number of informative CpGs in sample N1 dropped from 1200 to 519 when one mismatch was allowed instead of three, and the mapping efficiency from 1971682 (5.5%) to 1465798 (4.1%). The result is that while 25% of the reads were lost (and 50% of the informative CpGs), the benefit is improved accuracy.

Figure 4.6 Effect of altering stringency of mapping. Genomic sequence is in red. The bottom panel shows which reads mapped to this region when three mismatches were allowed. The black arrows point to mismatched bases when comparing to genomic sequence, showing that these reads have inaccurately mapped here. When only one mismatch is allowed in the algorithm, the two reads which are arrowed in red now no longer match and so do not appear in the top panel, allowing for a more accurate assessment of methylation.



4.2.5.4 PCR bias

It was also noted that some reads all mapped to exactly the same positions with identical starts and stops. This occurred in particular regions and probably reflects preferential PCR amplification. This problem was resolved by collapsing all reads that had identical starts and stops, after confirming visually that these reads did indeed have identical sequences. The result is a more graded start/stop that would reflect the natural variation in starts/stops of fragments generated by sonication. Figure 4.7 illustrates the reads that mapped both before and after collapse- e.g. at positions 99930 or 99939.

4.2.5.5 Strand discrepancy

As the data that enters Seqmonk is stranded as OT (original top), CTOT (complementary to original top), OB (original bottom), and CTOB (complementary to original bottom), it is possible to visualise which strand has given rise to which result. OT and CTOT results are grouped as "top", and OB and CTOB as "bottom". As all the oligos were designed with the sequence of the top strand, it would be expected that all reads should come from the bottom strand, both the original bottom and complementary to the original bottom (arising from PCR amplification). However, although 95.2% of the reads originated from the bottom strand, 4.8% did originate from the top strand (Figure 4.8A and 4.8B). Furthermore, the reads that originated from the top strand were clustered. Of concern, analysis of CpGs that had data from both top and bottom strands revealed that the methylation results from these were not always congruent. Figure 4.8C shows the discrepancies. 59.6% of the CpGs that had results from both top and bottom strands showed no discrepancy. 10.5% showed minimal discrepancy (<20% difference between the top and the bottom), 17.5% showed significant discrepancy (20-50%) and 12.3% showed severe discrepancy (50-100%). Although discrepancies could reflect true hemimethylation, as there were so few reads from the top strand and it was not evident where these originated, it was decided to exclude these from further analysis. One possibility is that they originate from partially denatured strands. As illustrated in Figure 4.8D, if some fragments of double stranded DNA are only partially denatured

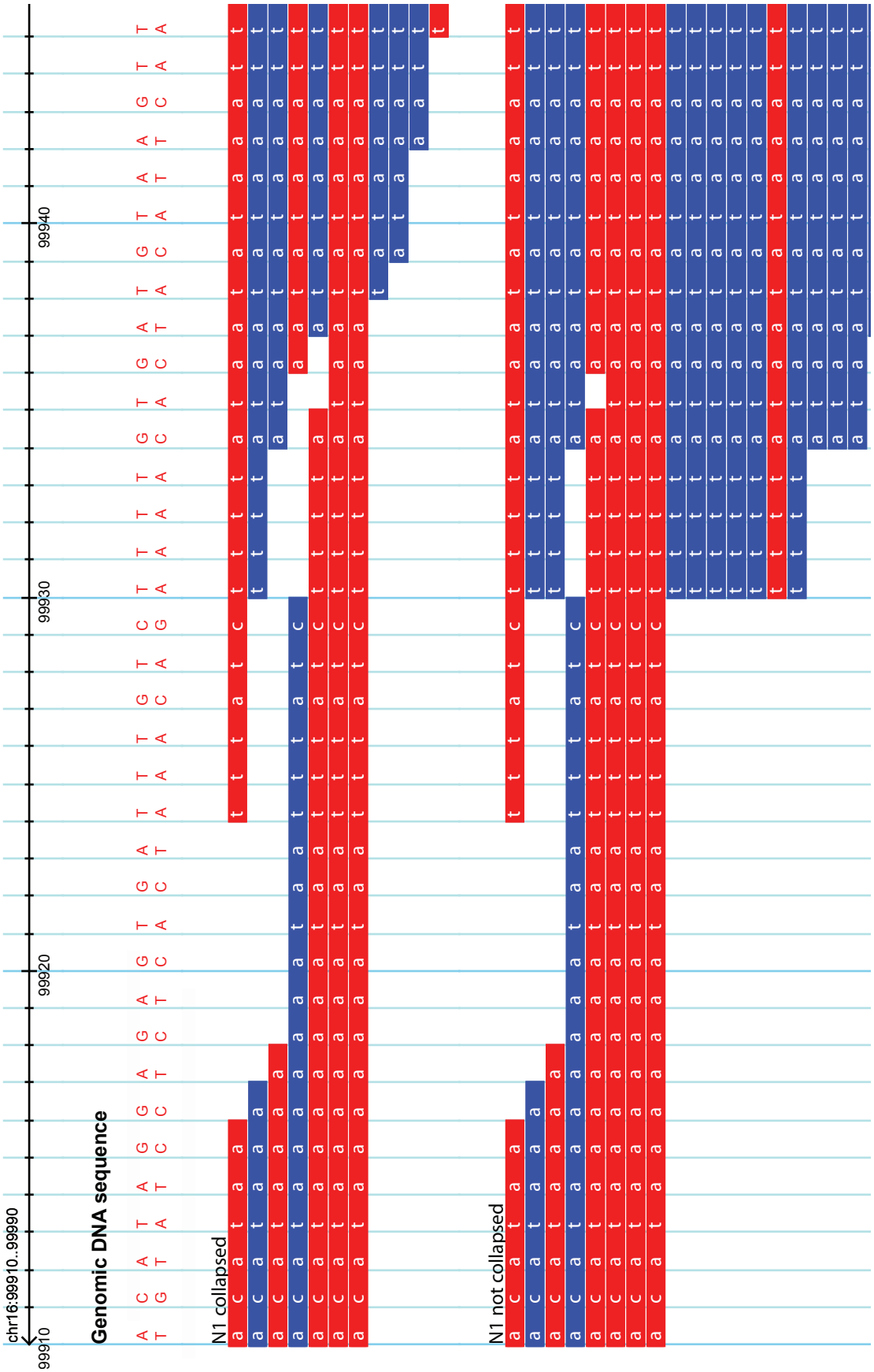
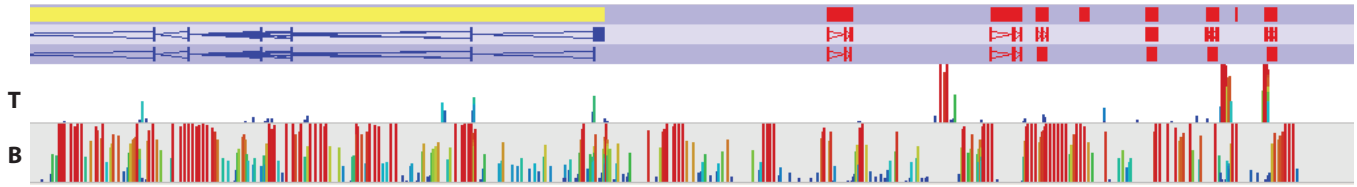
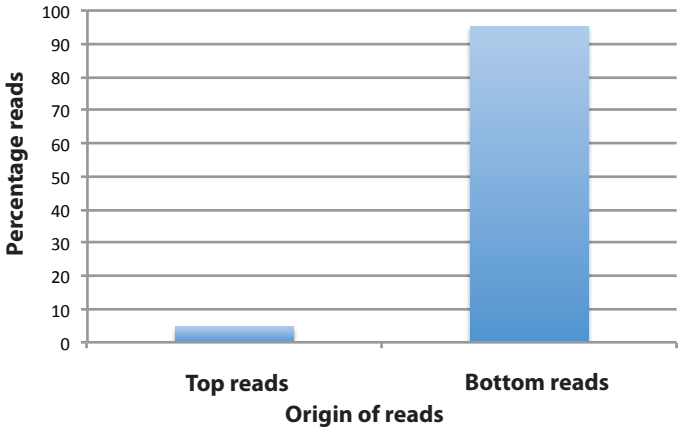


Figure 4.7 Effect of collapsing the reads. Some reads stack up with identical starts and stops, as seen in the bottom panel. This represents PCR bias. These reads are therefore collapsed to a single one as shown on the top panel.

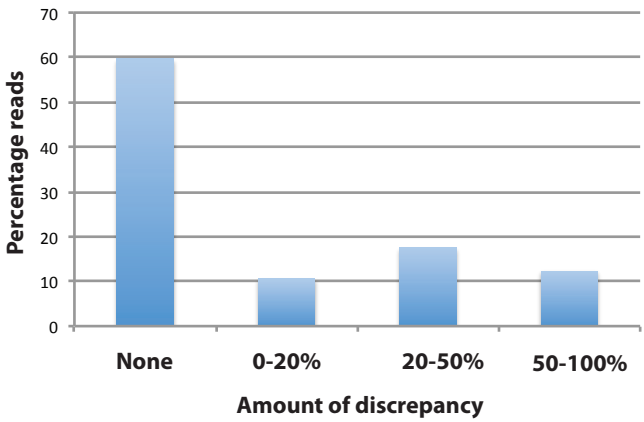
4.8A. Top vs bottom strands visualised on Seqmonk. T= TOP reads, B= BOTTOM reads



4.8B. Top vs bottom strands



4.8C. Amount of discrepancy



4.8D. Possible origin of top reads

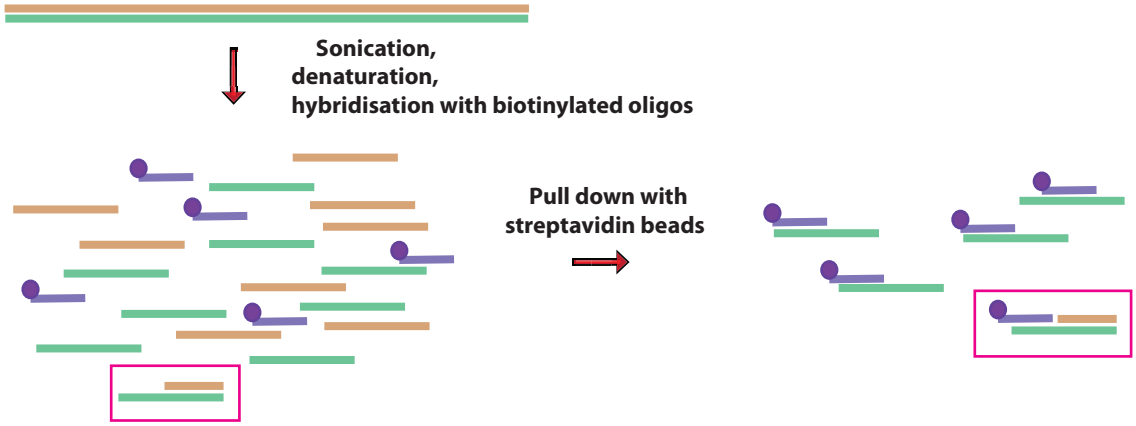


Figure 4.8 Strand discrepancy. **A**. Seqmonk track showing in the upper panel the reads coming from the top strand and in the lower panel the reads coming from the bottom strand. Note the reads from the top strand are clustered. **B**. Quantitation of the number of reads coming from the top vs bottom strands, showing fewer than 5% of reads are from the top strand. **C**. For the CpGs where both top and bottom reads occur, the % methylation calculated from the top reads was compared to that calculated from the bottom reads. While the majority of the results were within 15% of each other, some CpGs had significant discrepancy between the two results, with 12% of CpGs showing greater than 50% difference between top and bottom results. **D**. Cartoon representation of a possible source for the top reads, which may be due to inadequate denaturation of some DNA fragments as shown in the pink box.

(see pink box), this would allow some of the top strand to be pulled down with the bottom when the bottom hybridises to an oligo. In order to minimise this, the length of the denaturation step was increased although this had no obvious effects (the two samples that were processed after changing the protocol did not have fewer reads coming from the top strand, namely 7% for CLL2 and 10% for N2). It is possible that some of the top strands come from non-specific binding, however this would not explain the clustering of the top strand reads.

4.2.5.6 CpG density bias

Whereas 41% of all CpGs in the capture area are in CpG islands, only 8% of the 302 CpGs in the overlapping dataset are in CpG islands. CpGs from CGIs are therefore vastly underrepresented in our dataset (observed to expected ratio, O:E=0.2). The reasons for this may be lack of capture of these fragments, or their loss in the mapping process due to lack of unambiguous mapping. For the same reason that top strand fragments could be captured due to incomplete denaturation of high G/C content DNA, complete lack of denaturation and therefore no capture from very high GC content areas such as CGIs might be expected. Indeed, CpGs from which no reads were generated were clustered at certain CGIs.

Furthermore, an element of the lack of data from CGIs likely stems from the fact that these regions of high GC content make a poor sequencing template, which could bias against them at this stage of the protocol rather than at the capture stage.

Finally, these regions may be lost in the mapping process. Shorter reads from CGIs are likely to map to multiple areas due to the poor complexity of CpG-rich sequences, and therefore be discarded, which would lead to them being underrepresented. Whether the cause is poor capture, poor sequencing, loss during mapping or a combination of these explanations, it is clear that CpGs from CGIs are depleted in this dataset. Figure S4.2 illustrates the paucity of data from CpG rich regions.

In conclusion, after the initial results, several changes were made to the protocol and mapping/data analysis:

1. To minimise poor ligation of adapters, an extra clean-up step was introduced prior to the library preparation.
2. To improve mapping, number of PCR cycles during RPA was reduced from 32 to 30.
3. Sequencing was done to 50nt so that all samples could be compared.
4. All mapping was done single end so that all samples could be compared.
5. Bismark mapping was done with one mismatch allowed to reduce incorrectly mapped sequences.
6. PCR bias was minimised by collapsing reads that had identical start and finish positions.
7. The denaturation step was increased and the hybridisation carried out on a shaking platform to reduce the number of reads coming from the top strand.

The next two samples (N2, CLL2) were carried out with this new protocol.

4.2.6 Validation of the final protocol

Once the protocol and data analysis pipelines were optimised, it was possible to validate the results by comparing them with those from conventional bisulfite sequencing (CBS). Two of the same samples which were analysed by CBS and presented in Chapter 3 were also used for TBS (targeted bisulfite sequencing), namely N1 and CLL1.

Table 4.1 shows the characteristics of all the samples analysed by TBS.

	N1	CLL1	N2	CLL2
No. reads sequenced	36084225	27209965	63605278	51448788
No. reads mapped unambiguously	1465798	1588984	9030488	8246521
Mapping efficiency	4.1%	5.8%	14.2%	16%
No. reads with no alignment	32886074	22942957	51233425	40255672
No. reads did not map uniquely (%)	1732353 (4.8%)	2678024 (9.8%)	3341365 (5.2%)	2946595 (8.7%)
No. reads in target region	15788	22450	35606	36422
% mapped in target	.01%	0.014%	.004%	.004%
Enrichment*	3454-fold	5263-fold	9064-fold	8446-fold
Target region coverage (>1 read)	1231/2267 54%	1411/2267 62%	1773/2267 78%	1649/2267 73%
Target region coverage (>10 reads)	517/2267 23%	759/2267 33%	1024/2267 45%	648/2267 29%
Median read depth at target CpGs	8	11	23	23

- Enrichment= (reads mapped to target region/reads mapped to genome)/(size of target region/size of mappable genome)

Table 4.1. Characteristics of the samples analysed.

4.2.6.1 Correlations between Conventional and Targeted Bisulfite Sequencing

As shown in Figure 4.9A, for sample N1, 291 CpGs were available for analysis from CBS and 519 from TBS, with an overlap of 60 CpGs. Pearson's correlation of the methylation levels of these CpGs was 0.82. For CLL1, 303 CpGs were analysed by CBS and 760 by TBS, with an overlap of 78 CpGs. The Pearson's correlation for these was 0.91. These correlations are very high, particularly in view of the fact that the CpGs assessed by CBS were scored using a semi-quantitative method that binned them as 0/25/50/75/100% methylation as explained in section 3.3.2.

4.2.6.2 Examination of individual areas

Several individual areas are chosen to highlight the reproducibility of the results: Me16, which is a region which has high levels of methylation, while Me12 has low levels of methylation and Me22 has both high and low levels of methylation in a single fragment. These are illustrated in Figure 4.9C. It can be seen that there is excellent concordance between the TBS and CBS methods, lending credibility to this newly developed protocol. This is in agreement with the reported findings from capture bisulfite sequencing, where targeted high throughput bisulfite sequencing results were validated by traditional bisulfite sequencing (HODGES *et al.* 2009).

4.2.6.3 Identification of differentially methylated regions between normal and CLL

Ideally, the best test of TBS would be if it were able to identify some of the same hypomethylated CpGs between normal and CLL samples that had been reported in Chapter 3 by CBS. However, one of the shortcomings of these experiments is the low level of overlap between the two methods, which makes a detailed comparison difficult. Using TBS, there are 273 CpGs in common between N1 and CLL1, whereas in CBS there are 400 CpGs in common. However, between the 273 CpGs from TBS and the 400 from CBS, only 49 are in common between the two methods. Of these, 39 CpGs are not hypomethylated by either method, leaving only 10 possibly hypomethylated CpGs in common for analysis. One of these is hypomethylated only in the TBS group, six are hypomethylated by CBS

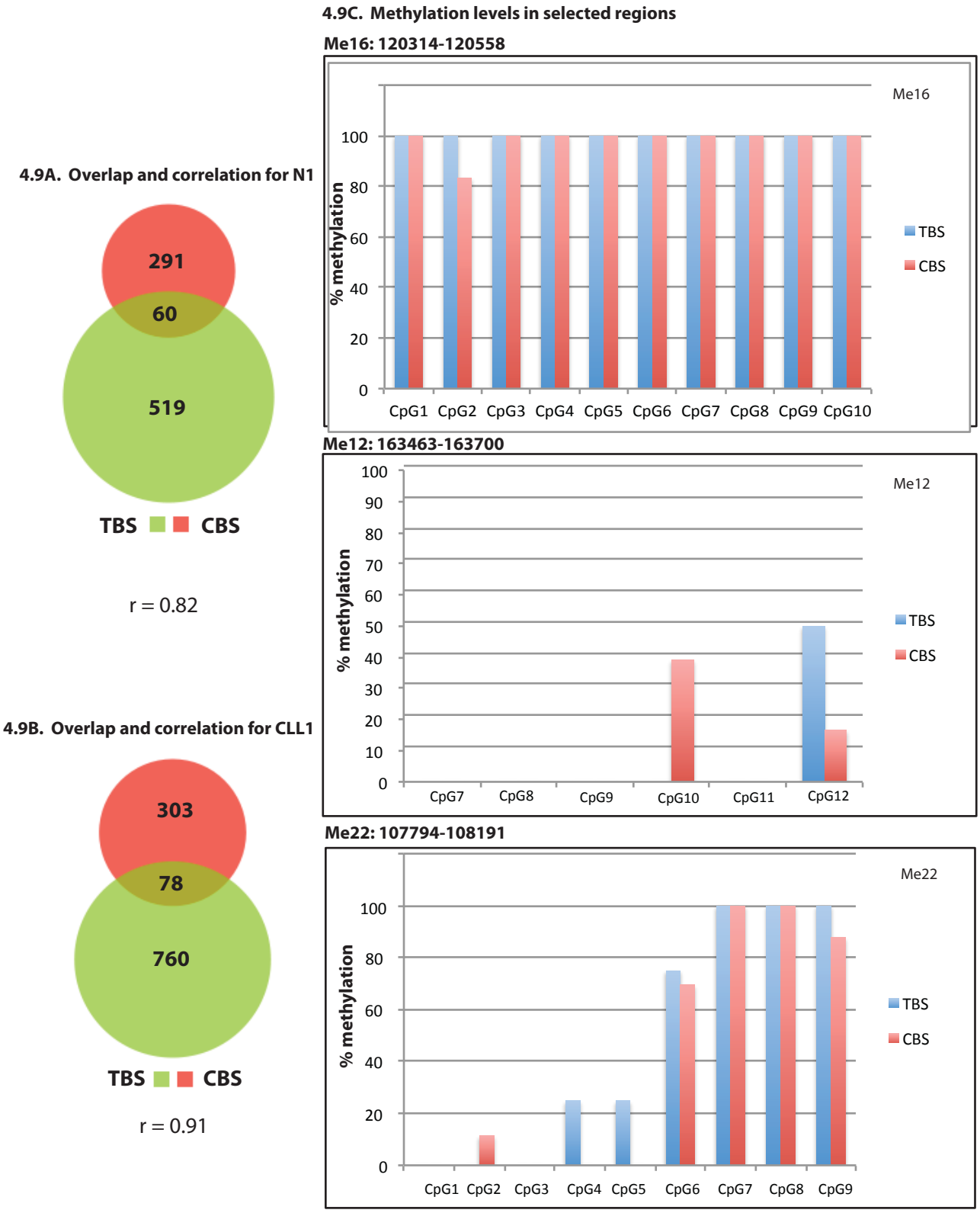


Figure 4.9 Validation of the technique. 4.9A. Venn diagram of the number of CpGs available for analysis by CBS (conventional bisulfite sequencing, red) and TBS (targeted bisulfite sequencing, green) as well as the CpGs in common to the two datasets for sample N1. **4.9B**. As above but for sample CLL1. **4.9C**. Three regions with individual CpG methylation levels shown for both CBS and TBS showing concordant results in areas of high, low and intermediate methylation. Methylation levels at individual CpGs in each amplicon is depicted in bar charts.

only and three are hypomethylated by both TBS and CBS. Although these numbers are small, it is reassuring that some of the same hypomethylated CpGs identified by CBS are also identified as such by TBS. The three hypomethylated CpGs in common are Me16 CpG2 (one of the statistically significantly hypomethylated CpGs identified by CBS, Chr16:120401, see Figure 3.12), Me24 CpG3 (Chr16:159158) and Me12 CpG12 (Chr16:163670).

The correlations between CBS and TBS on the same samples, examination of individual areas containing different levels of methylation, and the ability to identify some of the same differentially methylated CpGs between normal and CLL samples are excellent validation of CBS.

4.2.7 Results: Methylation changes in CLL patients

Once the protocol and data manipulation pipelines were optimised and CBS validated as a technique, two further CBS experiments were carried out, N2 and CLL2 in order to be able to pool two normal and two CLLs for detailed analysis of methylation differences between them.

4.2.7.1 Sequencing and mapping results

Table 4.1 (p. 163) summarises the sequencing and mapping reports for the four samples. The numbers of reads were 36M for N1, 27M for CLL1, 63M for N2 and 51M for CLL2. This near doubling of number of reads obtained between the first two and the last two samples reflects the fact that the output per flow cell has increased with time. The mapping efficiency (number of reads that map unambiguously) increased from 4.1% and 5.8% for N1 and CLL1 respectively to 14.2% and 16% for N2 and CLL2 respectively. As discussed previously, this may be due to reduction in "junk" DNA produced during PCR amplification. The percentage of reads that mapped to the target region was low (<0.02%) however this represented a significant coverage for this region, given its small size compared to the entire genome. Enrichment levels are calculated as =(reads mapped to target region/reads mapped to genome)/(size of target region/size of mappable genome). Enrichment levels were 3454-fold (N1), 5263-fold (CLL1), 9064-fold (CLL1) and 8446-fold (CLL2). Out of the 2267 CpGs in the target

region over half had at least one read covering them (54%, 62%, 78% and 73% for N1, CLL1, N2, CLL2). However, only CpGs which had at least 10 reads were used for further analysis, which corresponds to 23% for N1, 33% for CLL1, 45% for N2 and 29% for CLL2. The median number of reads at these CpGs was 8, 11, 23 and 23 for N1, CLL1, N2 and CLL2 respectively.

Figures S4.3, S4.4. S4.5, S4.6 show, for samples N1, CLL1, N2 and CLL2 respectively, a snapshot of Seqmonk containing methylation levels for all the CpGs across the target region, the number of reads for all CpGs and a histogram of the frequency of reads (number of CpGs which are covered by x number of reads).

Samples processed and analysed with the new protocol (N2 and CLL2) had more CpGs which were covered by a higher number of reads, which improves the reliability of the methylation calls for those CpGs. Ideally samples N1 and CLL1 would have been repeated with the new protocol in order to improve the quality of the data, however this was not possible due to time and cost constraints. Analysis of methylation differences between normal and CLL was carried out on the dataset containing CpGs for which there were at least 10 reads in all 4 experiments, namely a total of 302 CpGs.

4.2.7.2 Direction and magnitude of changes from normal to CLL

Figure 4.10A shows the distribution of methylation levels of these 302 common CpGs. The majority of the CpGs in both the normal and the CLL groups are highly methylated (85-100% level of methylation). However, there is a paucity of unmethylated (0-15% methylation level) CpGs. According to the widely accepted view that CpG methylation is bimodal, there should be a peak of unmethylated CpGs and a peak of highly methylated CpGs, at least in the normal sample. The relative lack of unmethylated CpGs most likely results from a sampling bias of the technique as discussed in section 4.2.5.6, since unmethylated CpGs are mostly found in CGIs and these are under-represented in this technique.

Fig 4.10A. Distribution of methylation levels of all CpGs

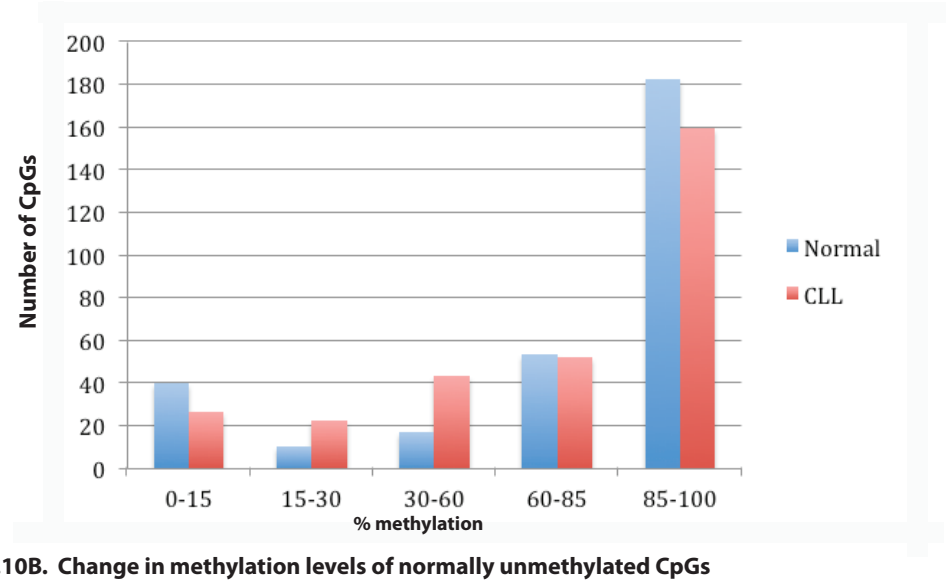


Fig 4.10B. Change in methylation levels of normally unmethylated CpGs

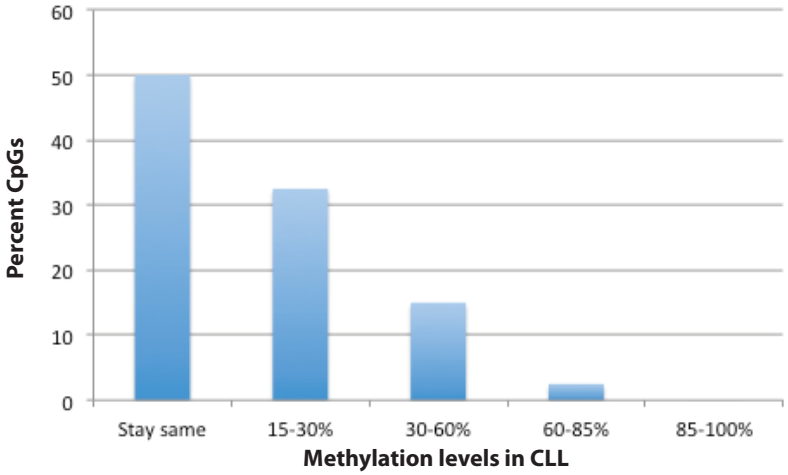


Fig 4.10C. Change in methylation levels of normally highly methylated CpGs

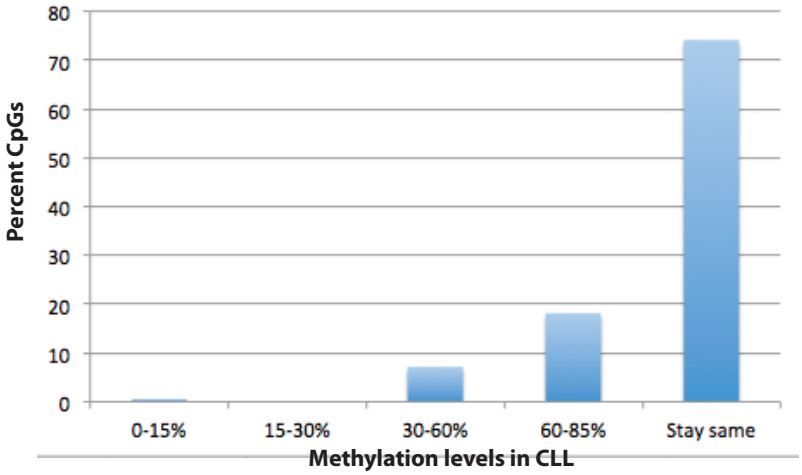


Figure 4.10 Magnitude and distribution of changes in methylation levels of CpGs between normal and CLL. **A**. Distribution of CpGs by methylation levels in normal (average of N1 and N2) and CLL (average of CLL1and CLL2). CLL samples have fewer highly methylated CpGs as well as fewer unmethylated CpGs. **B.** This graph follows the normally unmethylated CpGs and shows that while 50% are also unmethylated in CLL about 30% gain a little methylation to a 30-60% methylation level and few gain higher methylation. **C.**Highly methylated CpGs in normal samples are also highly methylated in CLL in over 70% of cases and those that lose methylation only partially lose it to 60-85% methylated, with very few CpGs losing methylation levels to below 60%.

From Figure 4.10A it can be seen that between normal individuals and CLL patients, there is a loss of both highly methylated and unmethylated CpGs and a gain in partially methylated CpGs, particularly CpGs methylated at 15-30% and 30-60% methylation levels. Figures 4.10B and 4.10C detail this transition further. Figure 4.10B looks at unmethylated CpGs in normal samples and charts their methylation levels in CLL. While half the unmethylated CpGs in normal samples were also unmethylated in CLL, the ones which gained methylation only partially gained it as 33% became 15-30% methylated and 15% became 30-60% methylated with only 2.5% of unmethylated CpGs gaining 60-85% methylation and none attaining full methylation. This is somewhat surprising as it might be expected that unmethylated CpGs, which tend to be in CGIs, would become highly methylated as they do in most malignancies. However as stated above, CpGs from CGIs were underrepresented, which may explain the lack of expected findings. Conversely, as shown in Figure 4.10C, for CpGs fully methylated (85-100%) in normal samples, 74% retained their methylation, 18% slightly lost methylation to a level of 60-85%, 7% to a level of 30-60% and <1% lost methylation to a level below 30%.

4.2.7.3 Distribution of changes

Figure 4.11 summarises the changes in methylation levels from normal to CLL. In the analysis in Chapter 3 on conventional bisulfite sequencing, "hypomethylation" was defined as a loss of ≥20% methylation and "hypermethylation" as gain of ≥20%. These numbers were chosen due to the semi-quantitative determination of methylation levels, which binned levels into 0/25/50/75/100%. With the TBS method, however, methylation levels are continuous and the concept of hyper- and hypomethylation could be refined. Currently there is no single accepted view in the literature as to what represents "hyper-" or "hypo"-methylation, or indeed what can be considered normal variation in methylation. When looking at DMRs in imprinted regions (which by definition should have 50% methylation) by pyrosequencing, these were found to be methylated at 35-65% (WOODFINE *et al.* 2011). For this reason it has been decided to allow a 15% variation to be normal, and called any greater deviation

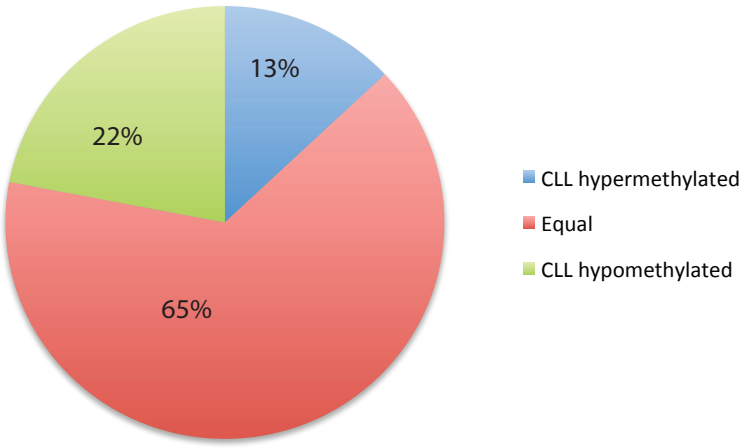


Figure 4.11A Summary of changes in methylation levels of CpGs between normal and CLL samples. Most CpGs are unchanged but 22% become hypomethylated in CLL and 13% hypermethylated.

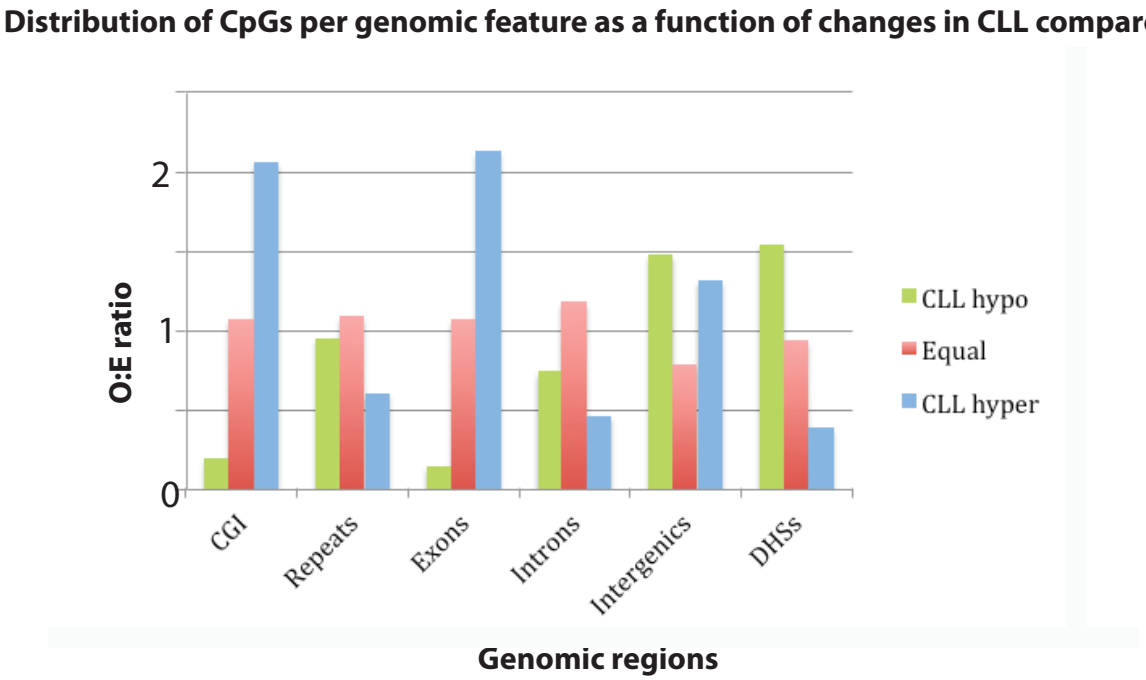


Figure 4.11B Distribution of CpGs per genomic feature as a function of changes in CLL compared to normal. The observed to expected ratio (O:E), is determined by dividing the frequency of distribution of CpGs in one category (hypo, hyper, or equal) by the frequency of distribution of all CpGs. Hypomethylated CpGs are more commonly found in intergenic regions and DNAse hypersensitivity sites, while hypermethylated CpGs are more commonly found in CpG islands and exons.

than that gain (hyper) or loss (hypo) of methylation. The 15% cutoff increases the sensitivity of finding methylation changes, but lowers its specificity.

Using this cut off, 65% of CpGs have unchanged methylation levels between normal and CLL (Figure 4.11A), 22% of the CpGs become hypomethylated and 13% hypermethylated. This is comparable to the results by CBS (62%/19%/12%), see Chapter 3.

Figure 4.11B shows the distribution of the changes in CpG methylation between normal and CLL according to genomic features. CpGs in CGIs are unlikely to become hypomethylated (O:E ratio = 0.2) and highly likely to become hypermethylated (O:E ratio = 2), which is expected as in the normal sample they would be completely unmethylated and in malignancy are known to be targets of hypermethylation. CpGs in repeats are less likely to become hypermethylated (O:E ratio = 0.6), which again is logical because they will be highly methylated already in the normal sample. However it is surprising that they are not particularly targets of hypomethylation in these samples. However this likely reflects the poor sampling of CpGs in repeats by TBS, as the reads corresponding to the repeats would not map back uniquely and would therefore be discarded. While 43% of all CpGs in the capture area are at repetitive elements, only 4.5% of the ones for which there is TBS data are in repeats, reflecting massive undersampling.

CpGs in exons are more likely to get hypermethylated (O:E ratio = 2.1) and less likely to get hypomethylated (O:E ratio = 0.15). CpGs in introns are less likely to be affected by either hyper- or hypomethylation (O:E ratio = 0.5 and 0.7 respectively). Conversely, CpGs in intergenic regions are more subject to change, both hyper- (O:E ratio = 1.3) and hypo- (O:E ratio = 1.5) methylation. Finally, CpGs in DHSs are targets for hypomethylation (O:E ratio = 1.5) but protected from hypermethylation (O:E ratio = 0.4).

4.2.7.4 Extreme changes in methylation

As well as looking at overall changes in methylation and where these CpGs may be located, it may be informative to look at extreme changes in methylation. For this reason, changes of more than two standard deviations from the mean in either direction in the distribution of changes in methylation between normal and CLL, were examined and labelled "extreme".

There were seven CpGs with methylation changes above 2SD from the mean, listed in Figure 4.12A, and 16 below 2SD (Figure 4.12B). Three of the hypermethylated CpGs surround HS-48, a DNAse hypersensitive site (DHS). This raises the possibility that all CpGs in that DHS are targets of hypermethylation. However, only one other CpG within HS-48 is also hypermethylated while the other two are roughly unchanged (5% and 16% loss in methylation, data not shown).

One of the 16 extremely hypomethylated CpGs is Me16 CpG2, which was already found by conventional bisulfite sequencing (CBS) to be statistically significantly hypomethylated in a range of CLL samples (see section 3.3.4.5, Figure 3.12).

4.2.7.5 Motif analysis

In trying to determine the mechanism for methylation changes in malignancy, one of the questions that arises apart from where the differentially methylated CpGs are located, is whether these CpGs occur within common sequences. This would suggest that the DNA sequence plays a part in targeting the CpGs for hyper- or hypomethylation. The way to identify this is by motif analysis of sequences surrounding hyper- or hypomethylated CpGs. This was conducted using the MEME software suite, on the extremely hyper- and hypomethylated CpGs.

4.2.7.5.1 Hypomethylated CpGs

Three motifs were returned when searching within 10bp of the hypomethylated CpGs. As shown on Figure 4.13, Motif 1 is [AG][AC]GG[TA]C[AG]G. Of the 16 extreme hypomethylated CpGs, 10 were found to have this motif. Three had

4.12A. List of CpGs extremely hypermethylated in CLL

CpG position	Location	Methylation normal	Methylation CLL	Difference methylation
94,625	5' HS-48; c16orf35 intron 8	33%	80%	-47%
95,226	3' HS-48; c16orf35 intron 8	56%	99%	-43%
95,229	3' HS-48; c16orf35 intron 8	24%	89%	-65%
151,377	"Me11", 5' HBZps	35%	96%	-59%
170,787	HBQ exon 2; CGI	0%	46%	-46%
170,807	HBQ exon 2; CGI	14%	73%	-59%
170,809	HBQ exon 2; CGI	25%	82%	-57%

4.12B. List of CpGs extremely hypomethylated in CLL

CpG position	Location	Methylation normal	Methylation CLL	Difference methylation	
96,787	HS-46; c16orf35 intron 8	100%	44%	56%	
100,124	c16orf35 intron 8	100%	54%	46%	
102,333	c16orf35 intron 7	98%	48%	50%	
108,788	c16orf35 intron 5	98%	52%	46%	
108,862	c16orf35 intron 5	93%	13%	80%	
120,401	c16orf35 intron 3	85%	35%	50%	**<-- Me16 CpG2**
133,088	HS-10	95%	48%	46%	
138,941	Intergenic region	98%	41%	58%	
142,469	HBZHSa	71%	20%	50%	
142,483	HBZHSa	73%	26%	47%	
142,566	HBZHSa	79%	28%	52%	
157,882	Intergenic region	51%	3%	48%	
158,239	5' HBA1ps	95%	35%	60%	
161,460	Intergenic region	96%	37%	59%	
165,667	Intergenic region	55%	7%	47%	
168,123	Intergenic region	81%	25%	56%	

Figure 4.12 A. List of the CpGs with extreme hypermethylation in CLL, as defined by being further than two standard deviations from the mean of differences in methylation between normal and CLL. **B.** List of the extremely hypomethylated CpGs in CLL. Note Me16.2, a CpG previously identified as consistently hypomethylated across a range of CLL samples using CBS, is found to be extremely hypomethylated using TBS.



Figure 4.13 Motifs associated with extremely hypomethylated CpGs in CLL

Motif 2, which is ATC[TA]GCCC and two had Motif 3, A[AC]AAAC. Motif 1 is most interesting both for its length (8bp; possible transcription factor binding site size) and for its frequency in the list. It is not found in any of the sequences containing hypermethylated CpGs. By comparison it is found in only one of the sequences of all the CpGs that are *not* hypomethylated. This gives an O:E ratio of 178 for its association with hypomethylated CpGs, a highly significant event that this motif is found in 10/16 hypomethylated CpGs ($p< 0.001$ by the binomial probabilities test).

The obvious question is what is able to bind to this sequence, and whether this could play a role in the hypomethylation of these CpGs. Using TOMTOM, an algorithm which predicts transcription factor binding to sequences, several known factors were identified as being able to bind Motif 1: t3R, PPARG, PPARA, NUR77, NR4A2, Dax1, COUPTF, ERR, FXR, and HNF4 alpha1 amongst others. To date, none of these factors have been reported in the literature to be elevated in CLL. However, a PPARG inhibitor has been found *in vitro* to induce apoptosis of CLL cells (INOUE *et al.* 2004), suggesting PPARG may be important in the survival of CLL cells. Furthermore, NUR77, a T-cell transcription factor, may play a role in T-CLLs (PEKARSKY *et al.* 2003).

4.2.7.5.2 Hypermethylated CpGs

Using the same criteria, the sequences with hypermethylated CpGs were searched. Three motifs were returned. As shown on Figure 4.14, Motif 1 C[CA]C[TA]C[GT][CG]GC[TC]GT[CT][CA] was identified in 5 of the 7 hypermethylated CpG sequences and none of the CpGs which were not hypermethylated. This may well be significant, but the motif identified is less interesting than the one identified at hypomethylated CpGs as its length is less likely to correspond to a DNA binding site factor. Motif 2 CGTGGA was found in 2 of the 7 hypermethylated sequences (and 6 of the 302 sequences for all CpGs, O:E ratio= 14, $p<0.01$). And Motif 3 [CT]TT[AC]CT in 2 of 7 hypermethylated sequences (vs 12/302; O:E ratio= 7, $p<0.05$).

MOTIF 1

Summary	
E-value	1.4e-003
Width	14
Sites	5
show more	

Sequence Logo

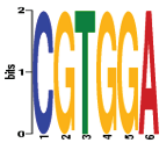


Name	Strand	Start	*p*-value	Sites
170809	+	3	1.55e-07	TAC CCCACGCGCTGTCC GCGCT
170807	+	5	1.55e-07	CCTAC CCCACGCGCTGTCC GCG
95226	+	3	1.95e-07	ACA CACTCTCGCCGTTC ATCTC
95229	+	0	1.95e-07	CACTCTCGCCGTTC ATCTCCAA
151377	-	1	8.87e-07	CTTCCTG CCCTCGGGCTGTCA C

MOTIF 2

Summary	
E-value	1.1e+003
Width	6
Sites	2
show more	

Sequence Logo

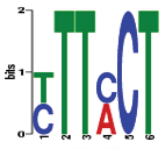


Name	Strand	Start	*p*-value	Sites
170787	+	2	3.38e-04	GC CGTGGA GCGCCTGGAC
94625	-	5	3.38e-04	GTAAAGTCCA CGTGGA AGTTA

MOTIF 3

Summary	
E-value	1.7e+003
Width	6
Sites	2
show more	

Sequence Logo



Name	Strand	Start	*p*-value	Sites
94625	+	16	6.66e-05	CCACGTGGAC TTTACT
151377	-	16	4.92e-04	CTTCCT GCCCTCGGGC

Figure 4.14 Motifs associated with extremely hypermethylated CpGs in CLL

Whether any of the individual factors which could bind to Motif 1 in the hypomethylated CpGs are involved in methylation changes in CLL is not yet known, but the observation that a particular sequence is associated with hypomethylated CpGs may provide an important clue to the mechanism behind methylation changes in malignancy.

4.3 Histone modifications

4.3.1 Rationale

As discussed in Chapter 1, DNA methylation and histone modifications are intricately linked, with some evidence of DNA methylation guiding the deposition of histone marks, as well as evidence to the reverse. Because DNA methylation anomalies in malignancy could well be secondary to alterations in histone modifications, it was important to assess these in CLL cells compared to normal. For example, it is known that H3K9me is enriched at repetitive sequences, which are deeply DNA methylated (Lehnertz et al., 2003), and that in some instances the deposition of the H3K9me3 comes before DNA methylation (STRUNNIKOVA *et al.* 2005). It would therefore be interesting to investigate whether this mark is lost in areas where DNA is hypomethylated in CLL or aberrantly deposited at sites of DNA hypermethylation. Conversely, activating marks may appear in CLL where DNA becomes hypomethylated, or be lost at sites of DNA hypermethylation.

4.3.2 ChIP-on-chip

At the beginning of the project, ChIP-on-chip was the routine method of assessing histone modifications. Customised PCR arrays covering 500kb on 16p were routinely used in our lab to provide good resolution visualization of DNA associated with most histone modifications. The aim for this aspect of the project was to profile active marks (H3K4me3, H3Ac, H4Ac) and repressive marks (H3K9me3 and H3K27me3) in CLL, with the aim of correlating these to DNA methylation changes if possible.

A total of 30 arrays were done. These arrays had previously been validated in the lab and extensively used. The samples analysed were normal whole blood, CLL cell lines and primary CLLs. The antibodies used were H3Ac, H4Ac, HeK9me3, H3K27me3, H3K4me3. All ChIP experiments were validated by RQ-PCR using standard probes routinely used in the lab and known to be positive and negative controls for each of these antibodies in normal lymphocytes and showed 20-100 fold enrichment in my samples. Labelling was checked by NanoDrop to ensure adequate concentration of the immunoprecipitated DNA and good dye incorporation, which ranged between 3.7-5.3pmol/µl. Despite this, hybridisation to the arrays was very poor. Expected peaks such as H3K4me3 at the promoters of housekeeping genes were on the whole either much smaller than those seen in previously analysed EBV-transformed lymphocytes, or completely absent. Missing peaks were often found to correspond to absent probes and some new peaks to hyperreactive probes on the arrays. This likely reflects a poor batch of arrays as all 30 used were synthesised in a single batch. Rather than try to validate what was evidently poor quality data by RQ-PCR, it was decided to carry out ChIP-Seq, which had by then become fairly routine and affordable.

4.3.3 ChIp-Seq

In deciding which antibodies to choose for ChIP-Seq analysis, known links between DNA methylation and histone modifications and the quality of the antibodies corresponding to these modifications had to be taken into account. Ideally one activating and one repressive mark would have been chosen. However, for the repressive mark, while H3K9me3 should have been a good choice because of its links to DNA methylation, from extensive experience with this antibody in our lab, it is a very poor one to use. Furthermore, as this modification is mostly present at repeats and sequencing reads that map to multiple sites (as repeats would) are discarded when mapping high throughput sequencing data, it was not a worthwhile antibody to use in ChIP-Seq. Another possibility for a repressive mark would have been H3K27me3 and although there have been some reports of links between H3K27me3 and DNA methylation (VIRE *et al.* 2006), H3K27me3 and DNA methylation are usually antagonistic

(MEISSNER *et al.* 2008) and anticorrelated (LYNCH *et al.* 2011). ChIP-Seq was therefore carried out on two activating marks, H3K4me3 and H3Ac, on the basis that they were good antibodies that work well for ChIP-Seq, which should maximise the usefulness of the data collected.

The ChIP-Seq was carried out as per the protocol described in Chapter 2. Two fresh primary CLL samples (CLL3, CLL4), and two CD19 cell samples (N3, N4) were used. The efficiency of immunoprecipitation was confirmed by RQ-PCR using known positive and negative control probes (see Appendix 1.B). The total amount of DNA following immunoprecipitation was estimated by Qubit and ranged from 50ng-300ng. Library preparation was carried out at the WTCHG. Due to cost considerations, not all samples were sequenced to the same depth. Therefore, for each antibody, one normal and one CLL sample were sequenced with one library and two lanes of sequencing, and for each of the second samples with the same antibodies, one library and one lane of sequencing were carried out. It must be stated that, while the samples on which ChIP-Seq was carried out were different from the samples on which TBS was carried out (N1, N2, CLL1, CLL2), the consistent localisation of methylation changes in CLL should allow a comparison between changes in histone modifications and changes in DNA methylation.

Figure 4.15 shows the tracks for H3Ac for the two normal and the 2 CLL samples and Figure 4.16 the tracks for H3K4me3 visualised on the UCSC broswer. These tracks cover the capture region. The green bars represent sites of extreme DNA hypermethylation (as identified by TBS, see Figure 4.12) and the red bars sites of extreme DNA hypomethylation. For H3Ac, there was nothing of striking interest in 16p, and certainly nothing related to the DNA methylation data. However, as can be seen on Figure 4.16, there is loss of H3K4me3 in both CLL samples at a site of maximal CLL DNA hypermethylation. This corresponds to the gene *HBQ*. Figure 4.17 shows the same site at two different magnifications. It can be seen that although there is indeed a peak at *HBQ* which disappears in CLL, the magnitude of that peak is not very high compared to, for example, that at *MPG* (red star). Nevertheless, while the peak at *HBQ* is smaller, it appears genuine

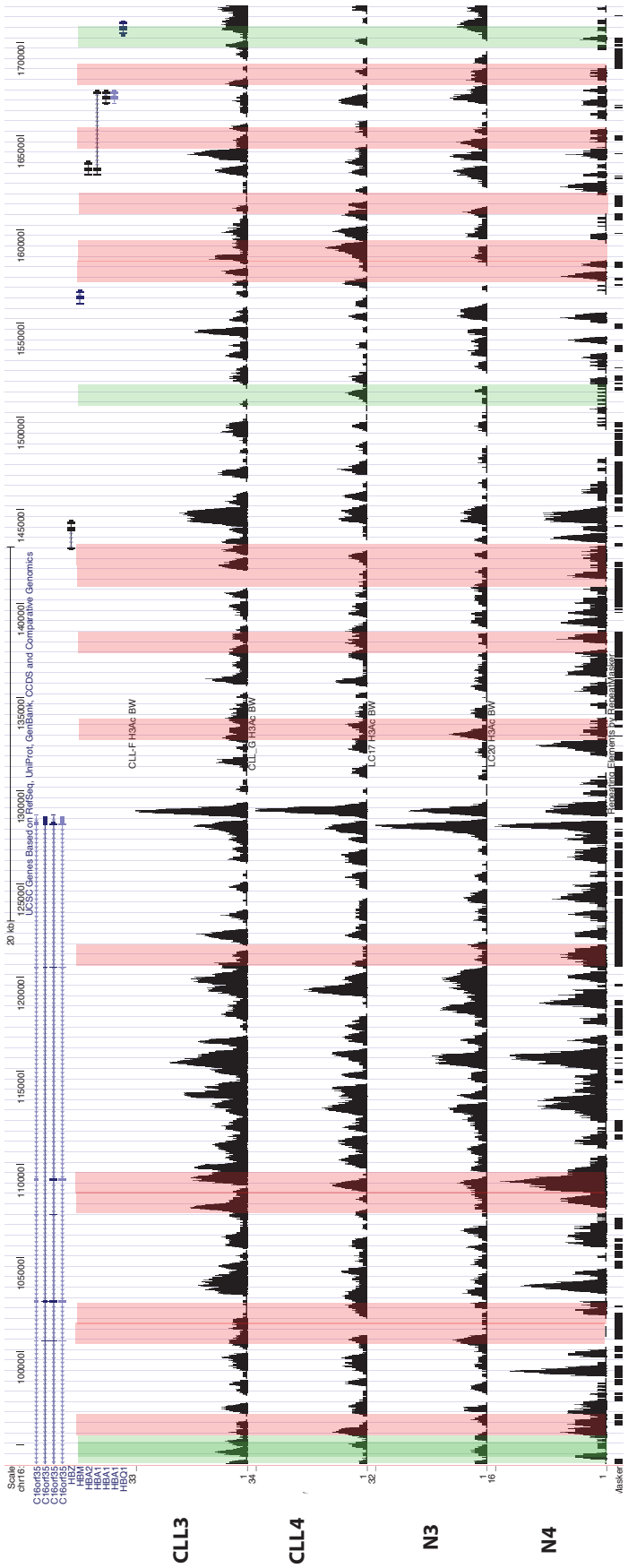


Figure 4.15 USCS view of two normal and two CLL tracks for the antibody H3Ac. The green bars highlight the sites of extreme DNA hypermethylation identified by TBS and the red bars sites of extreme hypomethylation in CLL. The top two tracks are CLL and the bottom two CD19 cells.

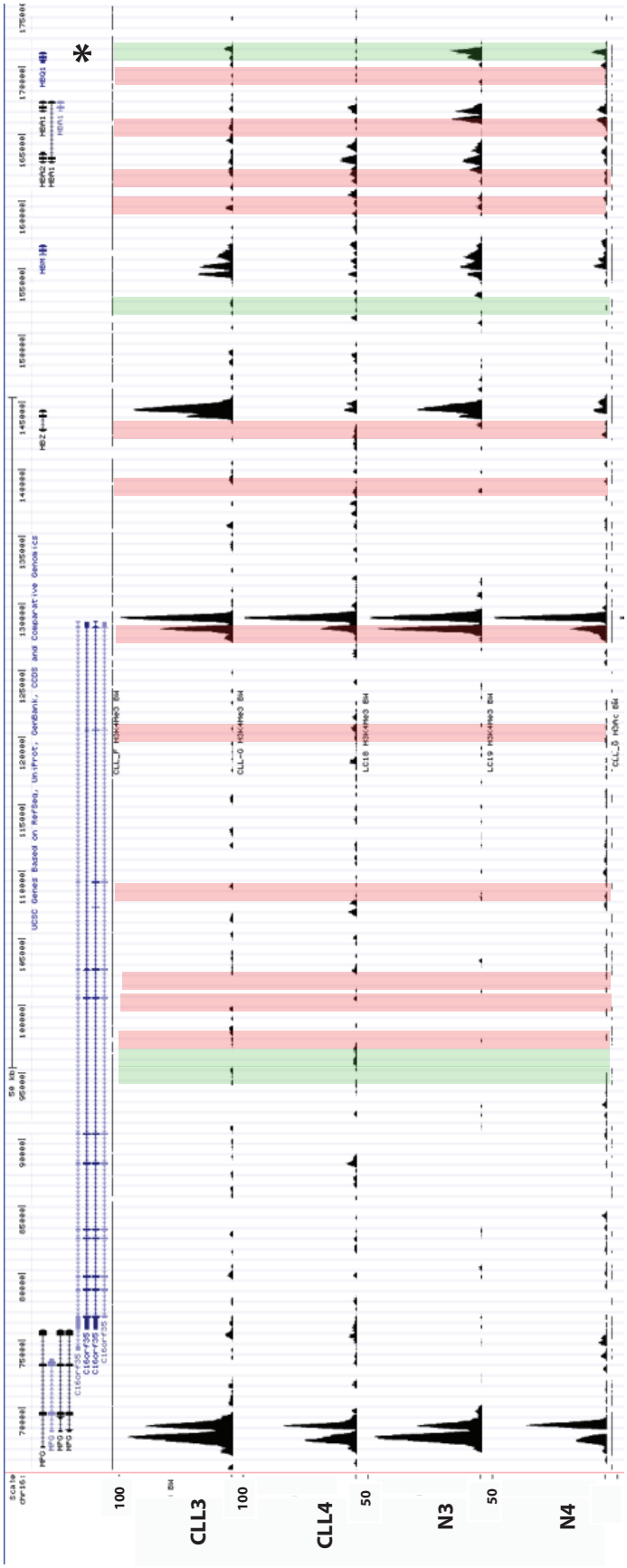
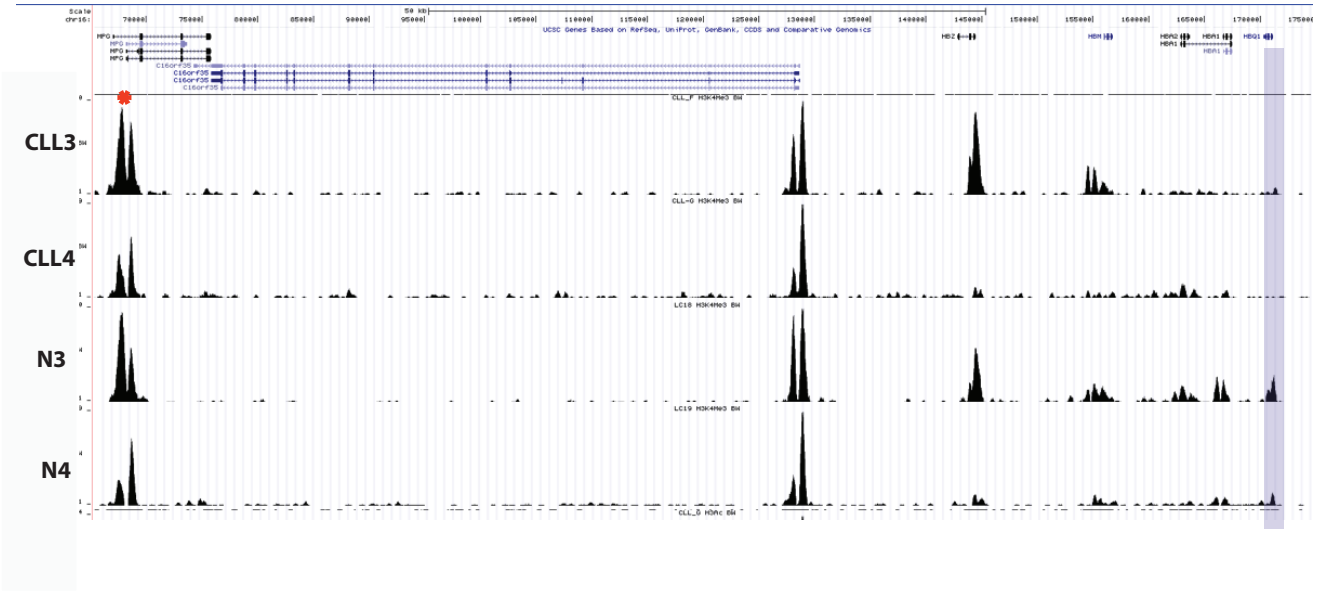


Figure 4.16 USCS view of two normal and two CLL tracks for the antibody H3K4me3. The green bars highlight the sites of extreme DNA hypermethylation identified by TBS and the red bars sites of extreme hypomethylation in CLL. The top two tracks are CLL and the bottom two CD19 cells. The starred green bar identifies a site of extreme DNA hypermethylation correponding to loss of the H3K4me3 peak in both CLL samples.

A



B

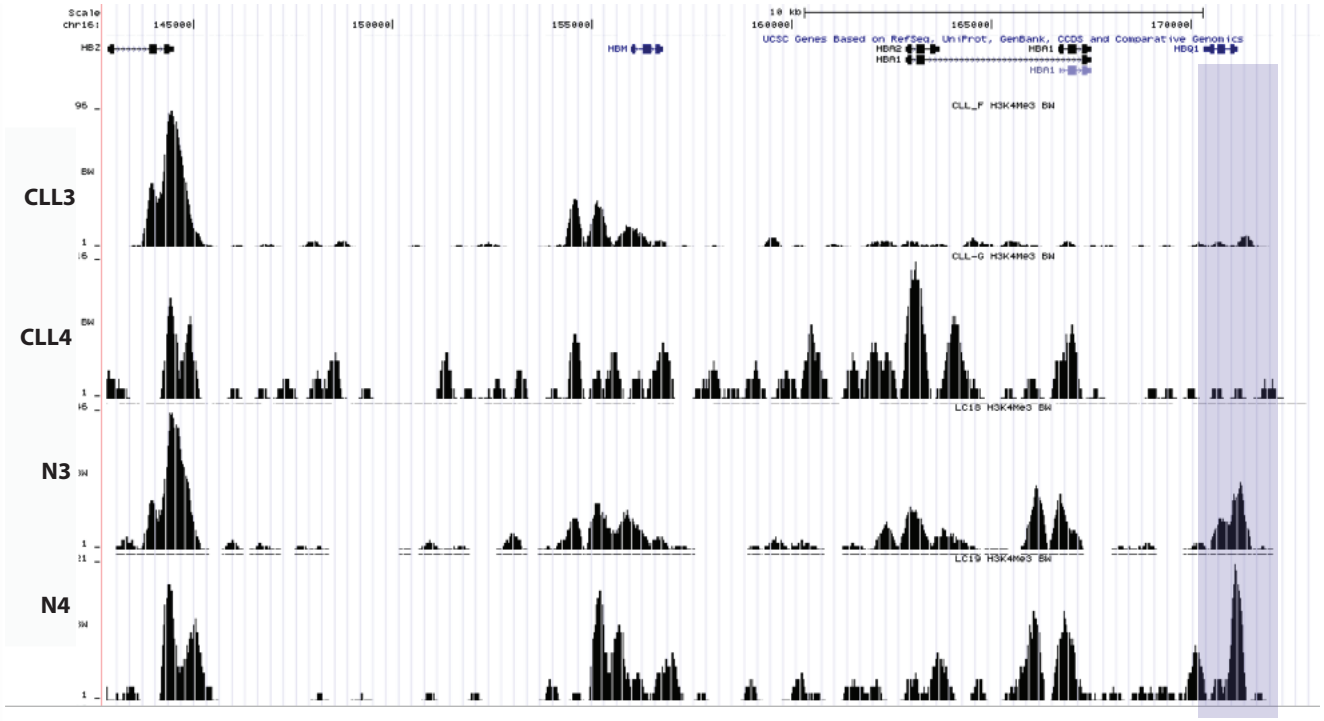


Figure 4.17 A. UCSC tracks for the two normal and two CLL samples for H3K4me3 peaks. The lavender bar highlights a peak that is present in both normal samples but absent in both CLL samples. This corresponds to a site of extreme DNA hypermethylation. The peak is seen at higher magnification in **B.**

due to its breadth. The question arises as to whether its loss in CLL at a site of maximal DNA hypermethylation is significant or not. It is known that He3K4me3 protects CpGs (usually in CGIs) from DNA methylation, so it is possible that in this case the primary defect could be loss of H3K4me3, thereby removing the protection from DNMTs and allowing these CpGs to become hypermethylated. Clearly this is not the case for all hypermethylated CpGs, as the other areas of extreme hypermethylation were not associated with such changes in histone modifications.

It is possible that there are several mechanisms to explain sites of DNA hypermethylation and that this is only one of them. Clearly, it is also possible that the co-localisation of a loss of H3K4me3 peak in both CLL samples and a site of DNA hypermethylation in CLL is coincidental. The question arises as to how frequently peaks of H3K4me3 disappear in both CLL samples compared to normal genome-wide. Of the 21686 peaks common to both CLL samples and the 25970 common to both normal samples, 64% of peaks are found in both normal and CLL samples. Differential peaks present in both normal samples and absent in both CLLs represent 13% (3769) of all peaks. However, there were no data on the DNA methylation status of CpGs in those areas. With time permitting, it would have been important and interesting to analyse a number of these loci by CBS to see whether and how frequently DNA methylation appears when H3K4me3 peaks disappear.

Conversely, at 6.2% sites in the genome (1805 peaks), H3K4me3 peaks were present in both CLLs but absent in both normal samples, however none were in chromosome 16p and there was no linked DNA methylation data with which to compare these.

While the bulk of the focus in the rest of this thesis is on DNA hypomethylation, the finding that DNA hypermethylation may be associated with sites of loss of H3K4me3 in CLL would be interesting to follow up in investigating a mechanism for DNA hypermethylation in malignancy.

4.4 Discussion

4.4.1 Technical approach

In developing targeted bisulfite sequencing (TBS), a period of optimization was required for the individual steps in the protocol and for the mapping and data analysis following high throughput sequencing. To reduce poor mapping efficiency, it was necessary to reduce the concentration of RPA primers, which reduced "junk" DNA produced from self-priming. The effect of PCR bias was reduced by collapsing reads with identical starts and stops and strand discrepancy was eliminated by using only bottom strand reads in the analysis. By sequencing the same two samples which had been sequenced by conventional bisulfite sequencing (CBS), it was possible to validate this technique, as correlations between CBS and TBS were excellent (0.82-0.91) and individual areas examined showed concordant results between the two techniques, with both techniques able to identify identical CpGs as extremely hypomethylated. Similarly to other published protocols, combining capture, bisulfite modification and high throughput sequencing is a useful technique for obtaining reliable methylation levels at individual CpGs.

In focusing on chromosome 16p, histone modifications in this region were also examined, as the cross-talk between DNA methylation and histone modifications is known to be bilateral, and to focus solely on DNA methylation may miss the primary defect causing its alteration in malignancy. As there are no published analyses of histone modifications in CLL, it was planned to thoroughly investigate a number of histone marks in CLL compared to normal. Both activating (H3Ac, H4Ac, and H3K4me3) and repressive marks (H3K9me3 and H3K27me3) were selected for analysis on PCR Sanger arrays specifically designed to cover the 16p region at high resolution. However, a poor batch of arrays precluded these from yielding meaningful results and the approach was changed to carry out ChIP-Seq, which was by then more of a routine and affordable experiment, and to do so in two normal and two CLL cases using two robust and reliable antibodies to assess H3Ac and H3K4me3.

4.4.2 Summary of epigenetic results at 16p

4.4.2.1 DNA methylation

A strong correlation of 0.85 in the methylation status of individual CpGs between the two CLL samples analysed by conventional bisulfite sequencing (Chapter 3) suggests that particular CpGs are targets for changes in methylation in this malignancy. This was confirmed, as particular CpGs which were hypomethylated in the initial two cases were statistically significantly hypomethylated when comparing 14 normal to 14 CLL cases. Although there is variability in methylation changes between CLL cases, the consistent targeting of particular CpGs suggests this is not a stochastic effect and the reproducibility of the sites affected suggests the mechanism may involve the DNA sequence, chromatin structure or accessibility of DNA.

Finding #1: Particular CpGs are consistent targets for hyper- and hypomethylation in CLL.

Both conventional bisulfite sequencing and targeted bisulfite sequencing revealed that, of the CpGs studied, most (62-65%) remained unchanged in their methylation status from normal to CLL. Of the CpGs that changed methylation status, slightly more became hypomethylated than hypermethylated (19-22% vs 12-13%). The near equal amount of hyper- and hypomethylation would suggest that the mechanism is not a simple one of abnormal levels of enzymes or substrate.

Finding #2: There are nearly equal amounts of hypermethylation and hypomethylation in CLL.

It was interesting to note that CpGs which were differentially methylated between normal and CLL samples followed a pattern of partial change only. Most normally unmethylated CpGs stayed unmethylated or gained a little methylation in CLL, while most fully methylated CpGs either stayed fully methylated or lost a little methylation in CLL. Because CpG methylation is a binary state, partial methylation reflects the situation in a population of cells

with a small drop in methylation at a particular CpG representing loss of methylation at that CpG in a fraction of the population. Partial demethylation would be compatible with a passive mechanism where, with time, the changes are gradually perpetuated and deepened. It would, however, also be possible that the mechanism for demethylation is in fact complete, as would be expected from an enzymatic process, and that the partial changes reflect a balance between complete demethylation and *de novo* remethylation of the same sites. Nevertheless, partial demethylation favours a passive non-enzymatic process.

Finding #3: Gain and loss of methylation at individual CpGs is partial on a population basis.

Strikingly, the genomic distribution of the CpGs that changed methylation status was not uniform. While there was a predilection in the distribution of the sites that change methylation status in malignancy, others seem protected from change. CpGs which became hypomethylated were less likely to be in exons, and more likely to be in intergenic regions, repetitive sequences and DNAse hypersensitive sites (DHSs). These represent genomic features likely to harbour transcription factor binding recognition sequences. Transcription factor binding sites are commonly found in intergenic regions, which contain numerous regulatory elements, whereas they are extremely rare in exons (VAN NIMWEGEN 2007), which seem protected from hypomethylation. Repetitive elements, such as *Alus*, also contain numerous transcription factor binding sites, a reason which has been given to partly explain their prevalence in our genome (COTNOIR-WHITE *et al.* 2011). Finally, clearly DHSs are sites of particular transcription factor binding. It has only very recently been reported that DHSs are enriched in hypomethylated regions in a variety of malignancies (HANSEN *et al.* 2011).

Finding #4: Hypomethylated CpGs are non-randomly distributed, being more present at repetitive elements, intergenic regions, introns and DHSs.

Regarding hypermethylation in CLL, this was found to be slightly more likely at CpGs that are in CGIs, unsurprisingly, as these are normally completely

unmethylated. Hypermethylation of CGI promoters is also a long established finding in malignancy, although it has also been shown that most differential methylation occurring in malignancy is not in CGIs themselves but in CGI "shores" within 2 kb of CGIs (IRIZARRY *et al.* 2009). The authors were not able to determine the significance of these differentially methylated shores, but did point to a functional role of these, as the methylation of these shores was closely linked to gene expression. In the current study, although 34% of the hypermethylation did occur in CGIs, the remainder occurred in other regions, such as exons. Non-CGI hypermethylation in malignancy is not a well-studied phenomenon, and its further investigation may provide clues towards the mechanism behind this occurrence. It is possible that the predilection for exons being involved in hypermethylation could be due to spreading of methylation from nearby hypermethylated CGIs, although this was not examined here.

Finding #5: As extensively reported in the literature, hypermethylated CpGs are non-randomly distributed, being more present in CGIs.

As particular CpGs seem to be targets for hyper- or hypomethylation in CLL, it was important to consider whether the DNA sequence surrounding these had common similarities such that it could explain how these CpGs are consistently modified. Motif analysis for sequences surrounding extremely hypermethylated or extremely hypomethylated CpGs revealed that there were indeed some DNA sequences that were enriched at these sites. In particular, one motif enriched at hypomethylated CpGs had a statistically significant O:E ratio of 178 as it was found in 10/16 sequences around hypomethylated CpGs but only one of the sequences surrounding CpGs not hypomethylated. Sequence motifs that are associated with normally hypermethylated CGIs have been described (SHEN *et al.* 2007) as well as motifs associated with sequences prone to and resistant to methylation in a cell line over-expressing DNMT1 (FELTUS *et al.* 2006) and in hypermethylated promoters in tumour cell lines (KESHET *et al.* 2006). They have not, however, been described at hypomethylated sites in malignancy. Although the putative transcription factors which could bind to the motifs described here are not known to be abnormal in CLL, the fact that particular short DNA

sequences are enriched at these sites certainly suggests the DNA sequence may bring specificity and targeting of the methylation changes.

Finding #6: Particular DNA sequence motifs seem to target CpGs for hyper- and hypomethylation.

4.4.2.3 Histone modifications

ChIP-Seq analysis of two normal and two CLL cases for two activating histone modifications showed the loss of an H3K4me3 peak in the two CLL samples, corresponding to a site of maximal gain in DNA methylation in CLL. As it is known that H3K4me3 marks cause exclusion of DNMT3L which maintains the associated CpGs unmethylated (OOI *et al.* 2007), it is possible that in this case, the primary defect is loss of H3K4me3, leading to lack of protection from DNA methylation. However, in the area studied, the converse did not hold true, as there was no gain of H3K4me3 where there was loss of DNA methylation. Genome-wide there were ~2000 examples of loss of H3K4me3 peaks from both CLL cases, but as there was no corresponding DNA methylation data with which to correlate this, the significance is uncertain. Nevertheless, this observation is worth investigating further at a future time. The limited study of histone modifications in CLL undertaken here does not exclude the possibility that the cause for DNA hypomethylation could well be a histone modification other than the two tested.

Finding #7: Loss of H3K4me3 occurs at a site of maximal DNA hypermethylation in CLL. Conversely, sites of hypomethylation are not associated with changes in H3K4me3 or H3Ac.

4.4.3 Elaboration of a hypothesis

From these findings, a possible hypothesis begins to emerge. One theory would be that the demethylation occurs in a passive manner, due to abnormal binding of transcription factors in malignancy. These are known to be dysregulated in malignancy, for example YY1 is upregulated in GI malignancies (CHINNAPPAN *et al.* 2009a; CHINNAPPAN *et al.* 2009b), SOX2 in osteosarcomas (BASU-ROY *et al.* 2011), DLX2 in melanoma (YILMAZ *et al.* 2011), GATA-4 in hepatoblastomas (SOINI *et al.*

2011), or chk1 in B-cell lymphomas (HOGLUND *et al.* 2011). These dysregulated transcription factors may bind in an uncontrolled manner in malignant cells. If this process occurs during DNA replication, and they obscure a methylated site on the parent strand, the daughter strand will escape methylation, resulting in a hemi-methylated CpG in the daughter cell. At the next generation, both daughter strands would be unmethylated. With increasing cell divisions, the CpG gradually loses methylation in the population, notwithstanding the fact that there could be re-methylation from the *de novo* methylation machinery. This hypothesis would explain all the current findings presented so far and is illustrated in schematic form in Figure 4.18.

At this stage, is not clear whether hyper- and hypomethylation in malignancy are mechanistically linked in any way, nor if there even is a single mechanism responsible for either phenomenon. In all likelihood, there may well be several mechanisms at play to explain certain divergent findings and while passive hypomethylation would be consistent with the findings presented above, this does not exclude the possibility of other mechanisms also being at play in the hypomethylation in malignancy. Nevertheless, this is the hypothesis that was then tested in the subsequent part of the project.

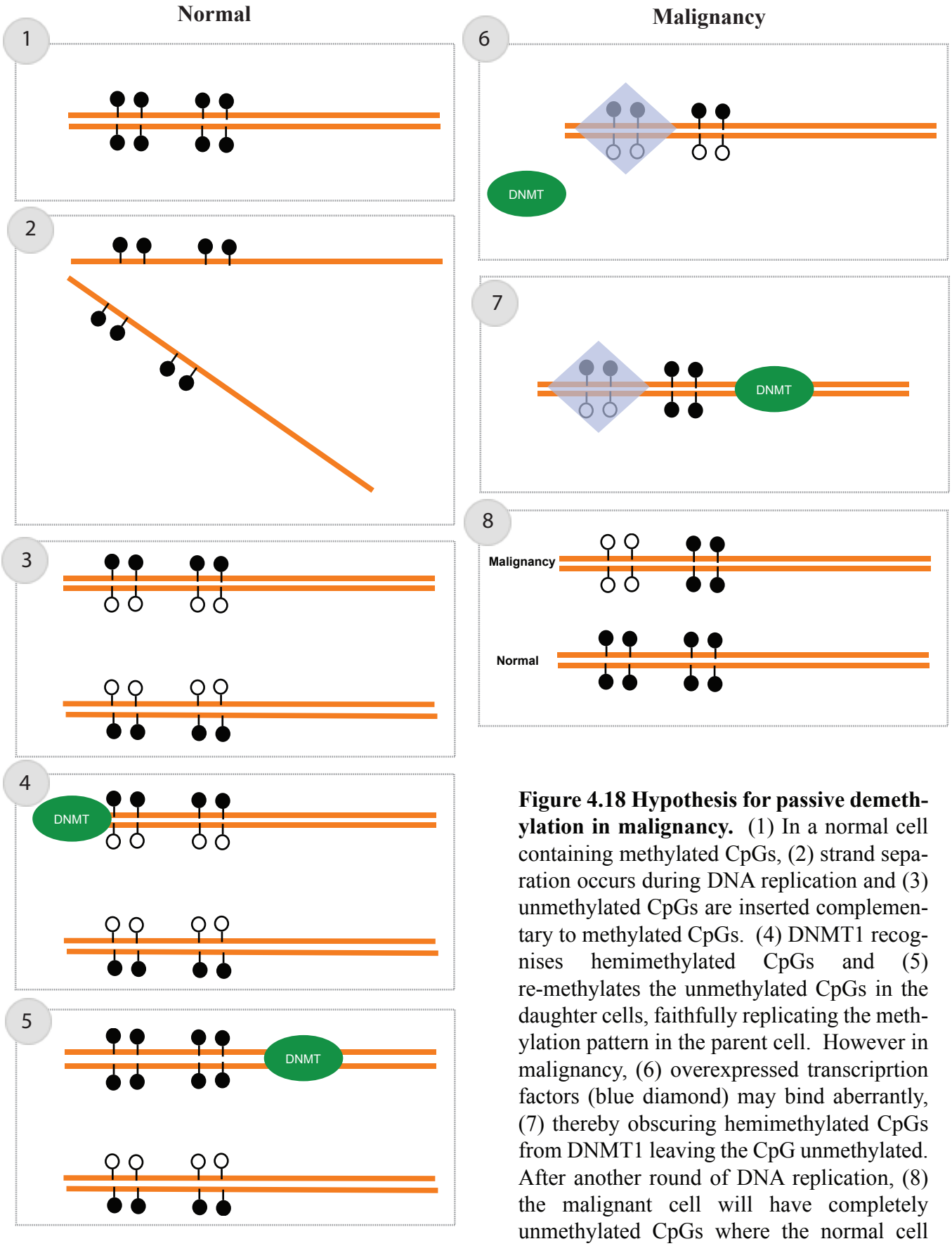


Figure 4.18 Hypothesis for passive demethylation in malignancy. (1) In a normal cell containing methylated CpGs, (2) strand separation occurs during DNA replication and (3) unmethylated CpGs are inserted complementary to methylated CpGs. (4) DNMT1 recognises hemimethylated CpGs and (5) re-methylates the unmethylated CpGs in the daughter cells, faithfully replicating the methylation pattern in the parent cell. However in malignancy, (6) overexpressed transcriprtion factors (blue diamond) may bind aberrantly, (7) thereby obscuring hemimethylated CpGs from DNMT1 leaving the CpG unmethylated. After another round of DNA replication, (8) the malignant cell will have completely unmethylated CpGs where the normal cell had completely methylated CpGs at the site of the transcription factor binding.

Chapter 5. Passive hypomethylation

5.1 Introduction

The hypothesis proposed here is that passive hypomethylation of specific CpG di-nucleotide sequences in malignancy results from the binding of aberrantly expressed transcription factors that obscure access of DNA methyltransferases which maintain methylation. If correct, several predictions can be made. Firstly, CpG sites that are very close together would become hypomethylated together, as they would be obscured by the same factor. Looking at direction of change (i.e. "hyper" or "hypo"- methylation) in adjacent CpGs as a function of the distance between them in the data presented earlier, CpGs that are within 2-10bp of each other frequently change methylation state together. This becomes less likely the larger the distance between them (Figure 5.1).

Another prediction from this model is that malignant cells would harbour some hemi-methylated CpGs, where the passive demethylation was in progress. Although hemi-methylation can be assessed by hairpin bisulfite sequencing (LAIRD *et al.* 2004a), this was not carried out here. However, there is evidence from the literature that ovarian carcinoma cells have significantly more runs of 2-4 consecutively hemimethylated CpGs than control somatic cells in sat2 alpha satellite DNA (SHAO *et al.* 2009). Such observations are consistent with the hypothesis proposed here.

Finally, we have some further evidence in our lab of this process occurring in the normal setting of differentiation which, while it was observed in a normal rather than malignant context, at least lends credibility to the hypothesis as a biologically plausible mechanism. During erythroid differentiation, a number of CpGs that lie in a regulatory region for *HBA* (called HS-40) become unmethylated as alpha globin becomes expressed. In early erythroid progenitors, where *HBA* is turned off and HS-40 is inactive and inaccessible, the CpGs are methylated. As the cells differentiate, HS-40 becomes a DNAse hypersensitive site, numerous

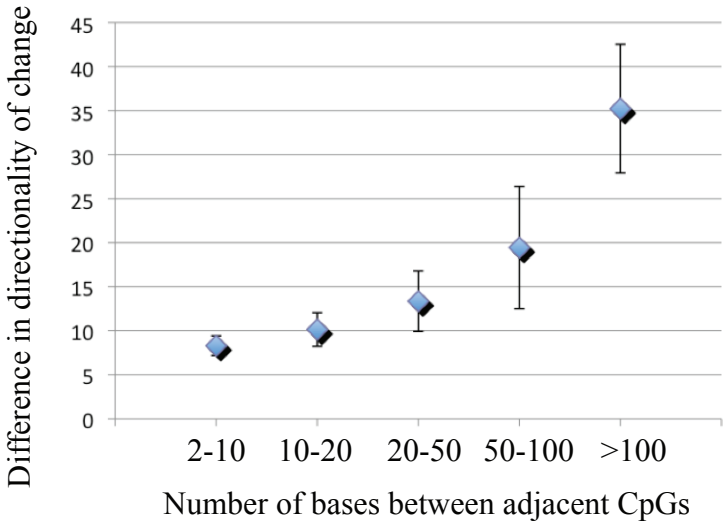


Figure 5.1 Change in CpG methylation between adjacent CpGs as a function of the distance between them. This was carried out on 330 CpGs for which there was bisulfite sequencing data in both normal and both CLL cases. The average % methylation of the two normal and two CLL cases were taken for each CpG. The absolute value of the difference between normal average and CLL average was recorded for each CpG. The difference in this absolute value between adjacent CpGs was then calculated. These values were then grouped based on the distance between the two adjacent CpGs. This shows that CpGs which are within 10bp of each other tend to behave in a similar fashion in CLL, either not changing status, or both becoming hypermethylated or both becoming hypomethylated. This likelihood decreases the greater the distance between adjacent CpGs. This would favour the implication of a locally acting factor on hyper- and hypomethylation.

transcription factors bind to it, and concurrently the CpGs become unmethylated (see Figure 5.2, data and figure from Dr Vip Viprakasit). The demethylation is not complete (as there is still a strong band of the higher molecular weight) and therefore unlikely to be an enzymatic process. Furthermore, the CpGs in question are not conserved in evolution, suggesting that the change in methylation does not play a regulatory role in *HBA* expression, but merely reflects the "footprint" of transcription factor binding and obscuring the CpGs from DNMT1 during DNA replication.

5.2 Preliminary studies

5.2.1 Enzyme levels

Prior to testing this hypothesis, it was necessary to ensure that, if there was passive hypomethylation, it does not arise merely from low levels of DNMT enzymes leading to global failure of maintenance methylation. The near equal amount of hyper- and hypomethylation in the CLL samples studied would argue against this, although one could envision the possibility of there being low levels of DNMT1 to explain loss of methylation in gene bodies and high levels of DNMT3A and/or 3B to explain hypermethylation at other sites such as promoters and CGIs.

Studies looking at levels of DNMTs in malignancy have found these to be increased. For example, DNMT1, DNMT3A and DNMT3B are increased in AML and Chronic Myeloid Leukaemia (MIZUNO *et al.* 2001), and DNMT1 alone is increased in cervical (SAWADA *et al.* 2007) and pancreatic cancer (KANAI and HIROHASHI 2007). A caveat is that normal levels of DNMT1 would not exclude the enzyme having a possible role in the loss of methylation. For example, low levels of DNMT1 may not be necessary in a prolonged manner, as transient depletion of DNMT1 has been shown to lead to stable demethylation and derepression of a transgene in melanoma cells (LORIOT *et al.* 2006). In that case, it would be more difficult to determine if this had been a transient event during malignant transformation. Equally, another possibility would be that there are normal levels of the DNMTs but that DNMT1 is not recruited to the replication fork. In the normal setting, both PCNA and UHRF1 are thought to play a role in recruiting

Fig 5.2A. Southern blotting of erythroid and non-erythroid cells at HS-40 (BglII/HpaII digest). The appearance of a smaller band in the erythroid cells reflects demethylation of the CpG.

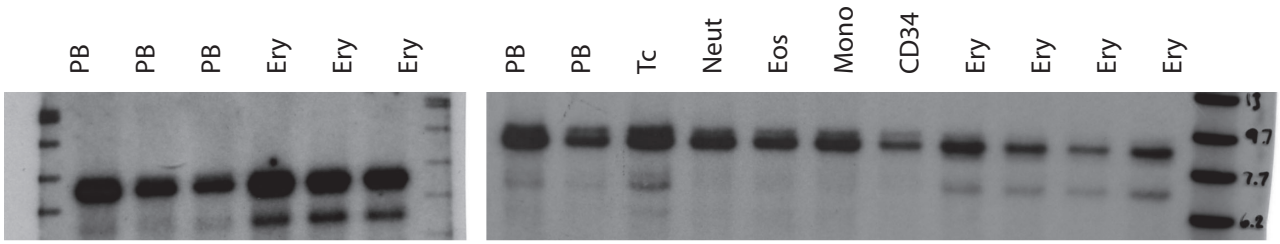
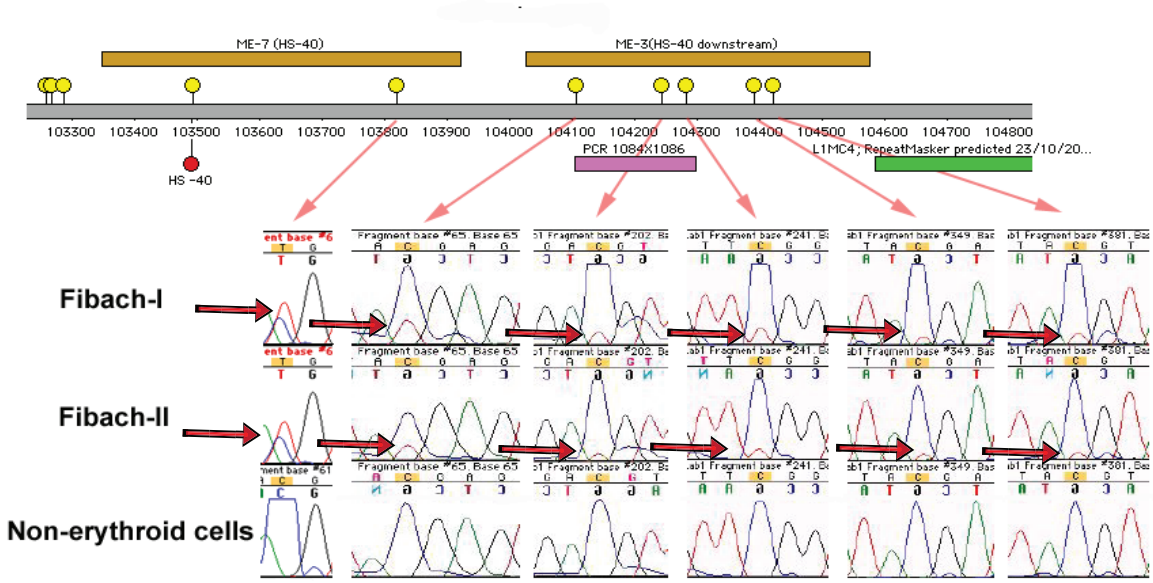


Fig 5.2B. Bisulfite sequencing of individual CpGs in HS-40 locus showing hypomethylation in erythroid cells, most pronounced at positions 103800 and 104100. Hypomethylation is evidenced by the appearance of a T peak in the sequencing.



Data courtesy of Vip Viprakasit, unpublished results.

DNMT1 to the replication fork ((BOSTICK *et al.* 2007; CHUANG *et al.* 1997) and decreased levels of PCNA and/or UHFR1 could explain passive hypomethylation due to failure of proper maintenance. Finally, while DNMT levels may be normal, this would not exclude their having a genetic mutation affecting their function, as is the case for DNMT3 mutations in AML (LEY *et al.* 2010).

Despite these caveats, it seemed reasonable to assess the levels of DNMT enzymes in the CLL samples under study. RNA was extracted from CLL and normal CD19 cells, and the expression levels of DNMT1, DNMT3A, DNMT3B, and DNMT3L, were compared by RQ-PCR on cDNA. The results are shown in Figure 5.3. There was no amplification with DNMT3B and DNMT3L for either the CD19 or the CLL samples. To exclude the possibility that this reflected failure of the expression probes themselves, RNA was extracted from human ES cells and a standard curve was calculated for both DNMT3B and DNMT3L. These were expressed in ES cells and the standard curves had R2 values of 0.99, confirming there was no problem with the expression probes themselves (data not shown). Therefore DNMT3B and DNMT3L were not expressed at detectable levels in CD19 cells or CLL cells. Expression levels of the other enzymes were calculated as ΔΔCT between the enzyme being investigated and 18S. Two or three CD19 samples were run on each RQ-PCR plate, and the results of the CD19 samples were averaged. A ratio between the ΔΔCT of enzyme:18S of the CLL samples was taken against the ΔΔCT for the average of the CD19 samples. DNMT1 and DNMT3A were normally or slightly overexpressed (2-4x) in the CLL samples (cell lines and primary samples), which are levels comparable to those previously found in malignancy. However one study has shown no change in DNMT levels in CLL (KN *et al.* 2004). In the present study, passive hypomethylation resulting from low levels of DNMT1 could be excluded.

5.2.2 5hmC levels

It has recently been suggested that 5hmC may play a role either in active or passive hypomethylation (see Introduction) but this has not been assessed in CLL. Although numerous techniques have been proposed to assess 5hmC

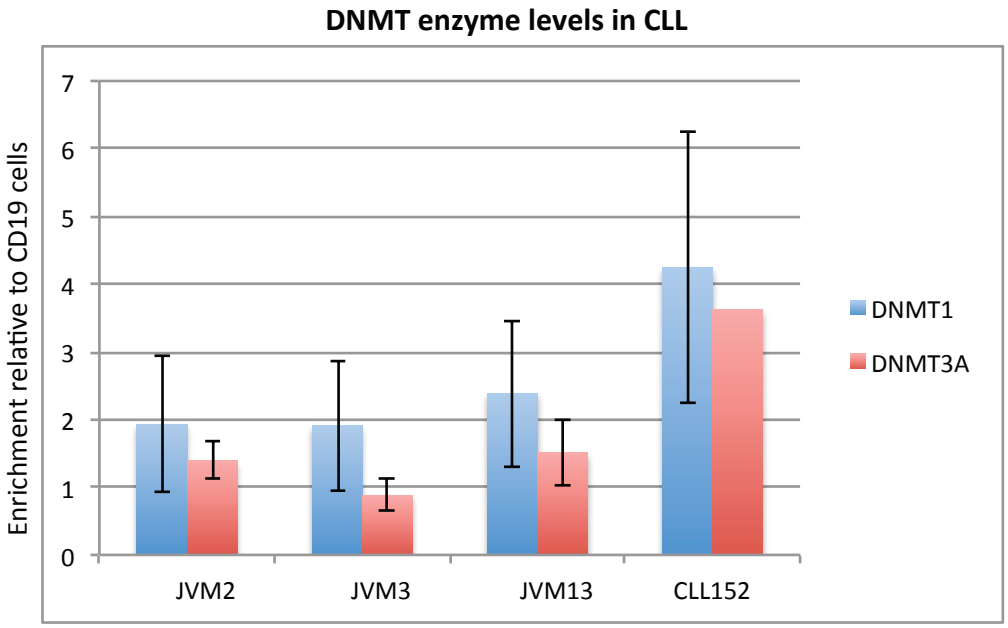


Figure 5.3 RQ-PCR plots of expression levels of DNMT1 and DNMT3A. Enrichment levels are calculated as ΔΔCT between the enzyme in question and 18S. This is then expressed relative to ΔΔCT enzyme:18S for CD19 cells. At least two CD19 samples were run per plate and the average of the CD19 samples was taken for comparison. Error bars represent standard deviations between separate RQ-PCR reactions done on identical cDNAs. JVM samples are CLL cell lines and CLL152 is a primary CLL.

(monoclonal antibody, JBP-1 enrichment, ELISA), it remains most straightforward to assess its total concentration in a sample by HPLC. Total genomic DNA is completely hydrolysed to single nucleotides and the sample subjected to HPLC. Quantitation of the sizes of peaks at known elution times determines the relative abundance of A, T, C, G, mC and 5hMC.

The abundance of these nucleotides was assessed in two CD19 and two CLL samples in a single HPLC run. Figure 5.4A shows the control run with the elution times of all the peaks in question. When the CD19 and CLL samples were run, there were no detectable 5hMC peaks in any of the samples. Figure 5.4B shows the quantitation for 5mC. It can be seen that the CLL samples had lower methylation levels than the normal samples, although the difference was small (2.7% CLL vs 3% CD19). The level of 5mC in CLL is the same as that reported in a previous study but there, normal samples (mix of B and T cells) had higher 5mC at 3.68% (WAHLFORS *et al.* 1992). There seems to be variability in HPLC estimations of 5mC: the median 5mC levels in leucocytes of human volunteers has been reported to be 4.1%, and the range was between 2.6-4.8% (KOK *et al.* 2007). The lack of any peak at the 5hmC elution time suggests that this is not a mechanism for demethylation in CLL.

5.2.3 Choice of area of study

Having ascertained that there did not seem to be a global explanation for the passive demethylation hypothesis, it was decided to focus the remainder of the study on the CpGs which are consistently hypomethylated in CLL, namely those discussed in section 3.3.4.5, which were statistically significantly hypomethylated when 14 normal and 14 CLL cases were compared (illustrated on Figure 3.12). From this list, five CpGs were chosen, namely Me13 CpG2, Me15 CpG7, Me16 CpG2, Me21 CpG5, and Me30 CpG2. These CpGs will be referred to as 13.2, 15.7, 16.2, 21.5, and 30.2 respectively. Figure 5.5A shows the mean methylation levels between normal and CLL for these five CpGs. It is noteworthy that Me16.2 is one of the CpGs which was identified as "extremely hypomethylated" by TBS in that it was the 8th most hypomethylated CpG of the whole dataset.

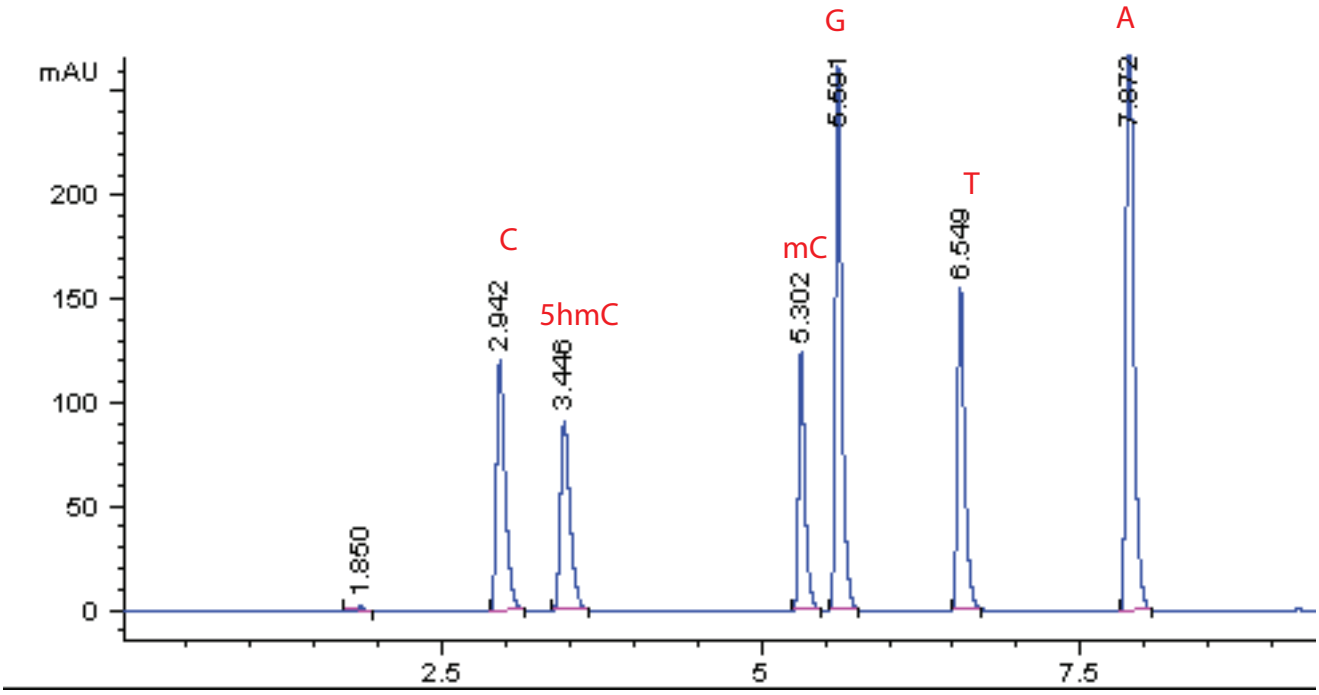


Figure 5.4A Control trace for HPLC showing the elution times of the different nucleotides.

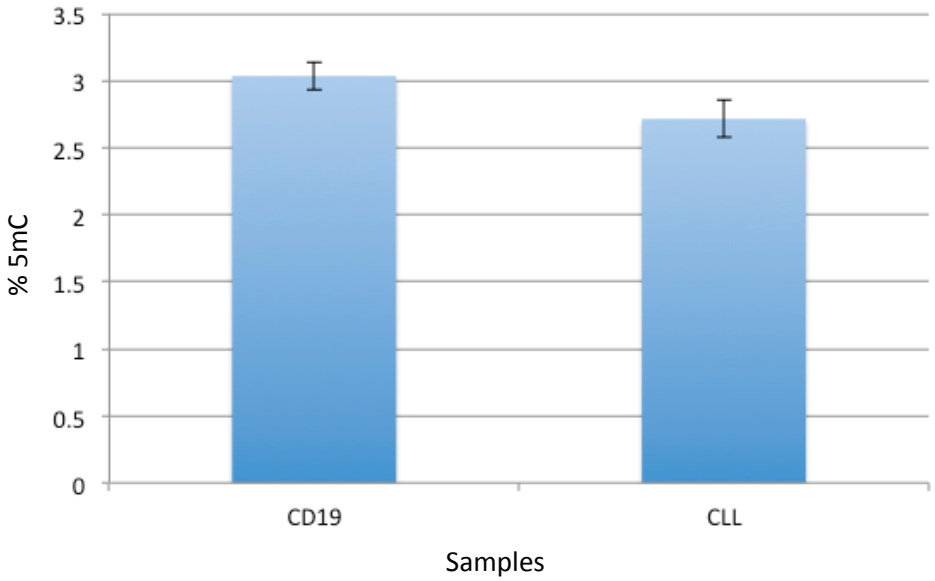


Figure 5.4B Quantitation of the 5mC peaks for normal and CLL samples. Error bars represent the standard deviation between two different samples analysed in a single HPLC run. The amount of 5mC in CLL is lower than in CD19 cells (2.7% vs 3%) however that is not as great a difference as previously reported.

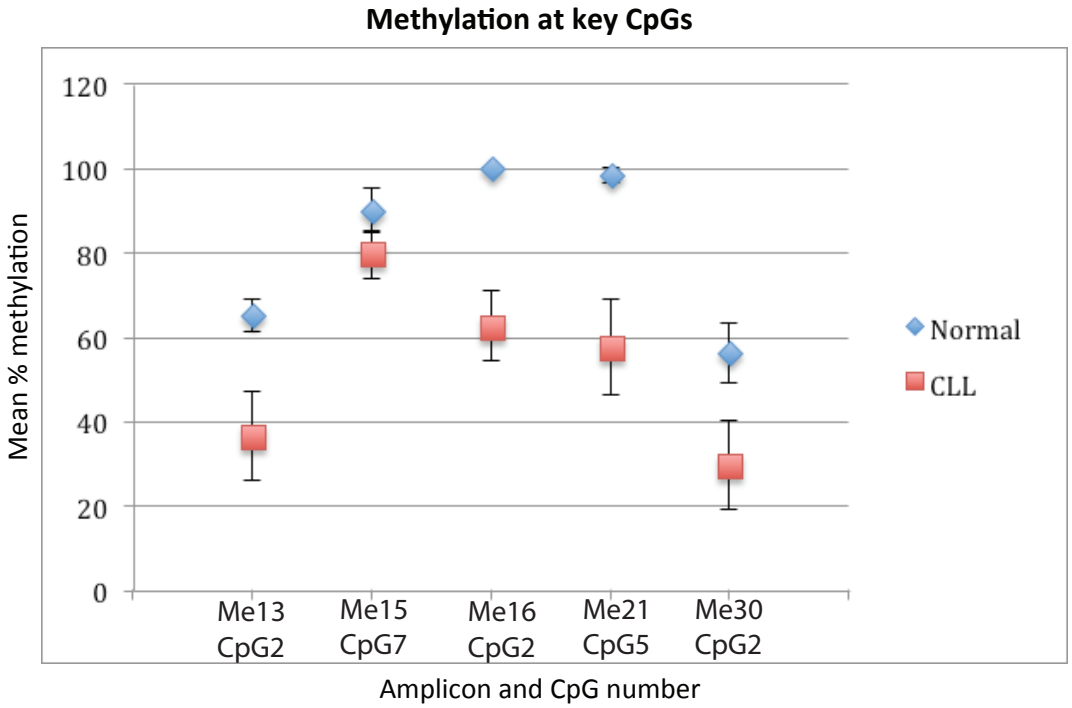


Figure 5.5A Composite graph of the five CpGs with consistent hypomethylation when analysed in 14 normal and 14 CLL cases. Blue diamonds indicate the mean methylation level of the CpG in the 14 normals and red squares in the 14 CLL cases. Error bars represent SD.

	Me13.2	Me15.7	Me16.2	Me21.5	Me30.2
Normal	65	90	100	98.2	56.25
JVM2	75	50	0	75	0
JVM3	0	75	100	75	0
JVM13	0	75	75	75	25

Figure 5.5B Table of the methylation level of individual CpGs. In orange are shown the normal values representing the average of 14 normal samples as on the above graph and in yellow are the values for the same CpGs in the three cell lines. Each cell line was sequenced independently at least three times.

As some of these investigations would need a lot of fresh material, it was decided to use the human CLL cell lines which are detailed in section 3.3.3.1 and which could be maintained in culture. It was therefore important to establish the pattern of methylation in these cell lines at these crucial CpGs. As shown in Figure 5.5B, for Me13.2, JVM2 was hypermethylated compared to normal but JVM3 and JVM13 were both hypomethylated at 0%. For Me15.7, all three cell lines were hypomethylated. For Me16.2, JVM2 was completely unmethylated, JVM3 was 100% methylated and JVM13 was partially demethylated at 75%. For Me21.5, all three cell lines were partially demethylated at 75% and for Me30.2 all three cell lines were demethylated, JVM2 and JVM3 the most at 0%. Although not every sample had equal methylation levels at each of these CpGs, having a range of demethylation could prove useful later in case it was possible to correlate level of demethylation and level of aberrant protein expressed.

5.2.4 Candidate proteins

Having identified the key CpGs on which to focus, there seemed to be some merit in first attempting a candidate protein approach, with the rationale that if proteins known to be over-expressed in CLL could bind to sequences surrounding extremely hypomethylated CpGs, these could be the proteins involved in passive hypomethylation. An initial search was made through the literature examining gene expression profiling in CLL, to see whether any reportedly overexpressed proteins in CLL had the capacity to bind any of the sequences surrounding the five consistently hypomethylated CpGs. Using TESS, a list of proteins that potentially bind these sequences from *in silico* predictions was compiled and is set out in a table in Figure 5.6.

There have been few studies of gene expression in CLL compared to normal. However, some studies have compared peripheral blood CLL to tonsillar B-cells (KLEIN *et al.* 2001), (ZHENG *et al.* 2002), or to CD19 cells (WANG *et al.* 2004). Others have compared bone marrow aspirates from CLL with those from healthy donors for transplantation (SONG *et al.* 2006). However, looking through these studies, none of the proteins predicted to bind to the sequences containing the hypomethylated oligos were found to be over-expressed in CLL.

13.2	15.2	16.2	21.5	30.2
Sp1	NF1	SRY	MyoD	myogenin
p300	AP-2	Pbx-1a	Myogenin	AP-4
myogenin	NF-BA1	C/EBP	AP-2	MyoD
CBF	IPF-1	FOXD3	GATA-1	Tal-1
TTK	MAF	HNF-1C	Sp1	Sp1
delta F1	LBP-1	NF-AT	Pbx-1a	LBP-1
E47		Elk-1	MAF	MZF-1
DBP			c-Fos	Myf-5
TCF-1			AML1a	
ADR-1				

Figure 5.6A Table showing a list of proteins with predicted binding to the oligos corresponding to each of these hypomethylated CpGs according to the TESS algorithm.

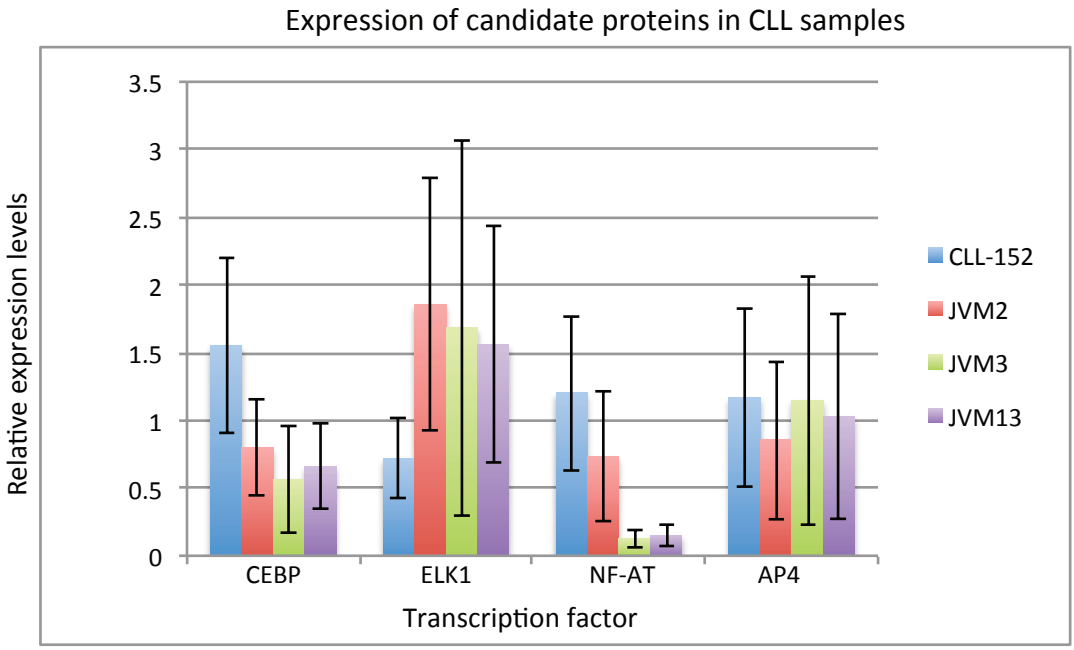


Figure 5.6B Graph showing relative expression of different transcription factors in the three cell lines and one primary CLL samples (CLL-152). For each sample, mRNA was extracted and converted to cDNA and then RQ-PCR carried out three times independently. Relative expression is determined from ΔΔCT between the expression probe and 18S and then expressed as a ratio between that in CLL and that in an average of two normal CD19 samples. Error bars represent SD (n=3). These are quite wide, which likely reflects the variability in the efficiency of 18S.

As an alternative method, it was decided to choose some of the proteins predicted to bind to the sequences surrounding the hypomethylated CpGs, as seen in Figure 5.6A, and assess whether any are over-expressed in the CLL cell lines. After investigation of the literature on the involvement of the transcription factors listed in Figure 5.6A in both the normal setting and in malignancy, the following transcription factors were chosen for analysis:

1. AP4. This is also known as TFAP4 or transcription factor AP4. It contains a basic helix-loop-helix domain. From its binding motif it is predicted that it could bind to 30.2. It has been shown to bind the immunoglobulin κ promoter (ARANBURU *et al.* 2001), and therefore could potentially be involved in CLL as this is a B-cell malignancy. It was also shown that in breast cancer cells, c-myc directly regulates the expression of AP4 and that AP4 was necessary for its tumorigenic activity (JUNG *et al.* 2008). Furthermore, AP4 is also over-expressed in colon cancer (GRATIO *et al.* 2009). It therefore seemed a reasonable candidate to consider, as it is implicated in both B-cells and malignancy.
2. NF-AT. This transcription factor, Nuclear Factor of Activated T cells, is predicted to be able to bind to 16.2. It is known to bind to DNA, albeit with weak affinity and often through other proteins (BOISE *et al.* 1993). However it has been shown to be transactivated in a proportion of CLL patients (SCHUH *et al.* 1996) which makes it a reasonable candidate to consider.
3. C/EBP-beta. This well-known transcription factor (CCAAT/enhancer binding protein) is predicted to bind to 21.5 and contains a basic region for DNA binding and a leucine zipper motif for dimerisation. It has been shown to bind DNA at the CR2 promoter in B-cells (CRUICKSHANK *et al.* 2009), making it a potential candidate for being involved in CLL. Furthermore, it has also been shown to have a role in malignancy as it regulates transcription factors critical for the proliferation and survival of cells from patients with multiple myeloma (PAL *et al.* 2009).
4. Elk-1. ETS-like transcription factor 1, a transcription activator, is predicted to bind to 16.2. It has been shown to be involved in breast

cancer tumour growth and survival (WYRZYKOWSKA *et al.* 2010). In the investigation of human leukaemias, it was shown that *MLL-AF4* rearrangements activate *Elk-1* through a Ras-dependent pathway (NG *et al.* 2010).

Figure 5.6B shows RQ-PCR data for the expression of these four transcription factors in the three CLL cell lines and one primary CLL sample (CLL-152), where expression is measured as ΔΔCT of the transcription factor to 18S cDNAs in the CLL relative to an average of two CD19 cDNAs. As can be seen on the graph, there is possibly some degree of over-expression of some of these factors, however the error bars showing standard deviation from three separate RQ-PCR reactions are quite wide and span the possibility of there being no over-expression in CLL compared to normal. These wide error bars may reflect the variability in amplification of the 18S expression probes as there was great batch-to-batch variability in the efficiency of RQ-PCR amplification of 18S. Due to these inconclusive results there was not enough evidence here to pursue any of these factors. It would however have been possible to follow certain lines of investigation. Firstly, it would have been possible to quantitate expression of these transcription factors to a different housekeeping gene. Secondly, it would have been possible to focus on a different set of candidate proteins, namely those predicted to bind the motifs identified by TBS as being associated with extreme hypomethylation. Finally, it would have been possible to carry out gene expression profiling of the CLL cell lines from which the DNA methylation data were obtained to see if different candidate proteins would emerge. However, it did not seem that these would be valuable uses of time and resources and instead, focus was made on *in vitro* assessment of potential candidate proteins from nuclear extracts.

5.3 Identification of proteins in nuclear extracts from EMSAs

The aim of this section of the project was to try to identify a DNA-binding protein that is over-expressed in CLL samples, using electrophoretic mobility shift assays (EMSA). EMSA allows the *in vitro* assessment of the interaction between

proteins present in a crude nuclear extract and oligos of a specified sequence, based on the observation that protein:DNA complexes migrate more slowly than free linear DNA fragments (the oligos) in a non-denaturing polyacrylamide gel electrophoresis.

In the ideal scenario, using this technique would lead to the identification of a band shift in the CLL samples that would not be present in the normal CD19 samples. The intensity of that band might correlate with the depth of hypomethylation in the samples. However, it was possible that no qualitative on/off phenomenon would be found. Rather, one might see a difference in the level of expression of a DNA binding protein. Furthermore, there may not be a correlation between the depth of demethylation and the level of expression of the protein if other factors influencing the binding differ between the samples (e.g. expression levels of associated proteins which could sequester the protein involved in demethylation).

5.3.1 Samples/protocol

5.3.1.1 CLL samples

For reasons discussed above, these experiments were mostly carried out with CLL cell lines. However, some fresh primary samples were also used. These were either stored overnight and the CLL cells isolated by lymphoprep or were taken fresh and isolated by CD19 beads as described for normal samples below.

5.3.1.2 Normal samples

CD19 cells were obtained for nuclear extract preparations. Initial attempts to isolate these from peripheral blood using MACS (magnetic-activated cell separation) were largely unsuccessful due to a combination of low number of cells obtained and of degraded proteins in the nuclear extracts once these were produced from the cells obtained that way. From 30-100ml of whole blood, 1-8x10^6 CD19 cells could routinely be isolated, which proved largely insufficient for downstream applications. In addition, even when sufficient numbers of cells were obtained, the process was so protracted and time-consuming (histopaque

isolation of PBMCs, MACS negative selection for removal of CD14 cells then MACS positive selection for CD19 cells) that the nuclear proteins had often degraded, as assessed on EMSA (data not shown). Therefore both the source of blood and the isolation technique were changed and CD19 cells were isolated directly from lymphocyte cones. These are available from the National Blood Service and are the by-product of the preparation of platelets from apheresis. They are supplied as 10-15ml but are highly concentrated for mononuclear cells to a higher level than buffy coats. Furthermore, the Dynabead system for isolation of CD19 cells was used, as this can be used directly from whole blood (or in this case, lymphocyte cones), cutting down several hours and manipulation of the samples, and thereby reducing the amount of protein degradation. From lymphocyte cones it was possible to routinely obtain 1-5x10^7 CD19 cells.

5.3.1.3 Nuclear extracts

As detailed in Chapter 2, initial attempts to prepare nuclear extracts with the traditional protocol according to Dignam *et al.* (DIGNAM *et al.* 1983) gave poor lysis (30-50%), whereas addition of 0.1% NP40 as in a different published protocol (OSBORN *et al.* 1989) consistently yielded 100% lysis.

5.3.1.4 Gel shift protocol

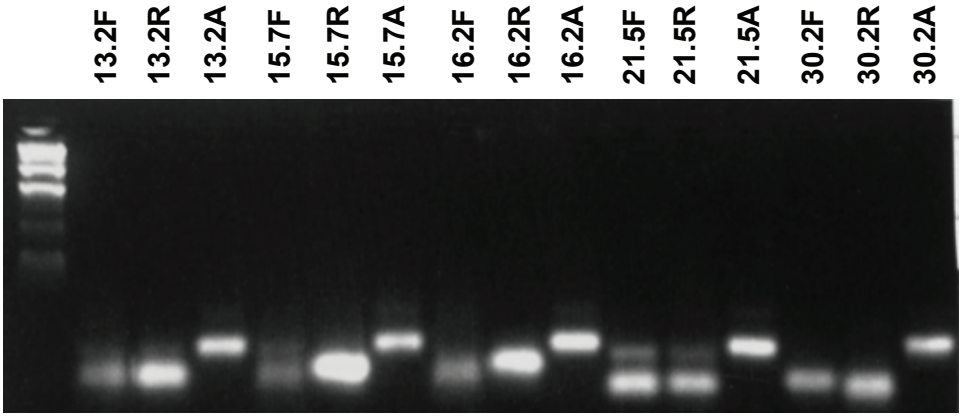
The EMSA protocol followed was the one published in *Molecular and Cellular Biology* (CROSSLEY *et al.* 1996) and is detailed in Chapter 2.

The oligos chosen for probes are shown on Figure 5.7A and are 30bp long, with 2bp at the 5'end of both the forward and reverse sequences being GG so as to be able to label with ^{32}P-dCTP. Figure 5.7B shows the annealed and unannealed probes.

Oligo name	Orientation	Sequences
Me13.2	Forward	5'GGAGCAGAGCACCTTG***CG***GCCTTACTCCTGCA3'
	Reverse	5'GGTGCAGGAGTAAGGCCGCAAGGTGCTCTGCT3'
Me15.5	Forward	5'GGGGCGTGGTGGCAGT***CG***CCTGTAGTCCCAGC3'
	Reverse	5'GGGCTGGGACTACAGGCGACTGCCACCACGCC3'
Me16.2	Forward	5'GGGAGCACCCACTAAA***CG***GAAAAAAACAAACA3'
	Reverse	5'GGTGTTTGTTTTTTTCCGTTTAGTGGGTGCTC3'
Me21.5	Forward	5'GGTGGGGCTGTGGCTG***CG***TTTCATCAACTGTC3'
	Reverse	5'GGGACAGTTGATGAAACGCAGCCACAGCCCCA3'
Me30.2	Forward	5'GGGTGCCAGGCATGGG***CG***TGGACAGCAGCTGG3'
	Reverse	5'GGCCAGCTGCTGTCCACGCCCATGCCTGGCAC3'

Figure 5.7A Table showing the sequences of the 5 oligos to be used as probes in EMSA. Highlighted in yellow are the CpGs which were identified as consistently hypomethylated. A sequence of 30bp was taken in total with the CG in the middle. Two Gs were added at the 5' end of both the forward and reverse sequences such that a total of four 32P-dCTPs would label each probe.

Figure 5.7B 1% Agarose gel of the annealed probes. F= forward, R= reverse, A= annealed.



5.3.2 Screening gel shifts

Once the oligos were annealed and radioactively labelled, gel shifts were carried out with all five of these probes. Probes 15.5 and 21.5 did not yield any band shifts. The probe 13.2 did yield two bands in the CLL cell lines as well as the CD19 cells. These are shown on Figure 5.8A and the bands are designated A and B. The intensity of the bands was quantitated. In order to correct for the amount of loading in the wells, control gel shifts were run alongside the test gel shifts, where an equal amount of nuclear extract was incubated with a probe for Sp1 binding. Band intensity for the test gel shifts was then calculated relative to the intensity of the band in the Sp1 gel (JVMband:JVMSp1/LCband:LCSp1). Quantitation showed that there was equal amount of band shift in the CLLs compared to normal for both bands A and B as seen on Figure 5.8B.

When 16.2 was used as a probe, 4 bands appeared, labelled A-D on Figure 5.9A. There were no qualitative differences; all four bands were present in the CLLs and the normal. However these bands were quantitatively more prominent in the CLLs. For example, the protein(s) associated with band D appeared to be present at 5-10x greater concentration in CLL than in the normals (Figure 5.9B). There also seemed to be some differential binding in CLL for bands A and C, albeit less than for band D. Interestingly, JVM2, which is the most hypomethylated CLL (0% methylation), had the highest level of the protein causing the band D shift.

Finally, the probe 30.2 did also show some increased band shifts in the CLLs, e.g. 1.6-4.5x for band A (Figure 5.10).

Although both 16.2 and 30.2 would have been possible candidates to investigate further, it seemed logical to focus on 16.2 as this was confirmed by TBS to be an "extremely hypomethylated CpG" whereas there was no TBS data for 30.2.

Fig 5.8

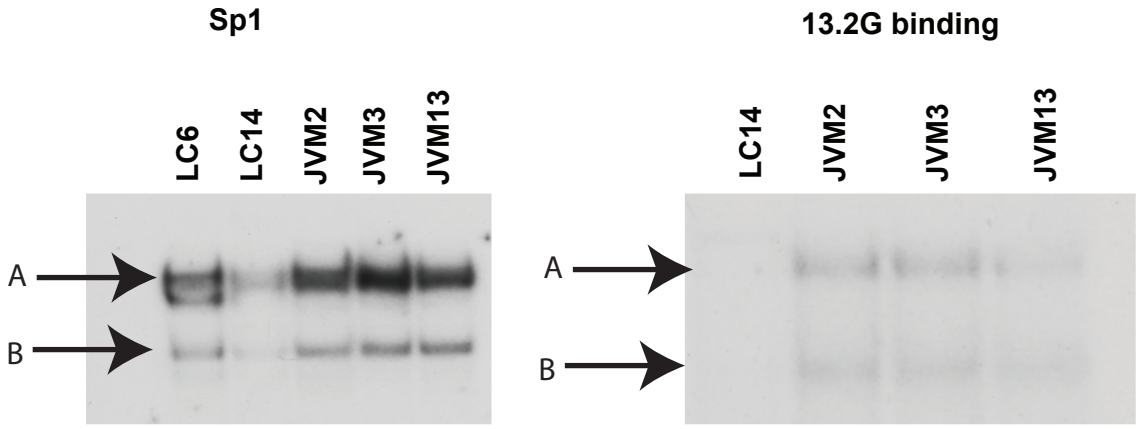


Figure 5.8A Autoradiographs of EMSAs carried out with two probes, Sp1 on the left and 13.2 on the right. Samples labeled LC are normal CD19 cells isolated from lymphocyte cones. JVM samples are CLL cell lines. The bands shifted on the EMSA are labeled A and B for both oligos.

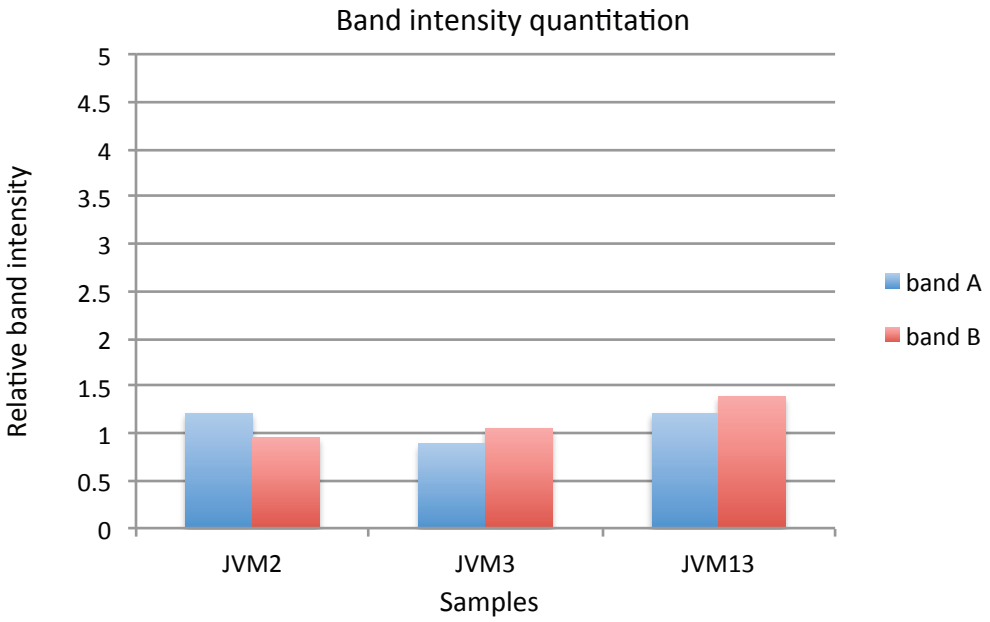


Figure 5.8B Quantitation from phosphorimager of the different bands. Each band is quantitated and then expressed relative to band B in Sp1. The ratio of oligo band: Sp1 band for each CLL sample is then compared to the ratio of oligo band:Sp1 band for the normal sample, giving the relative band intensity. Both bands A and B for 13.2 have similar band intensities to the normal samples, indicating that the proteins causing the band shifts to A and B are expressed in comparable quantities in the three CLL samples and the normal samples.

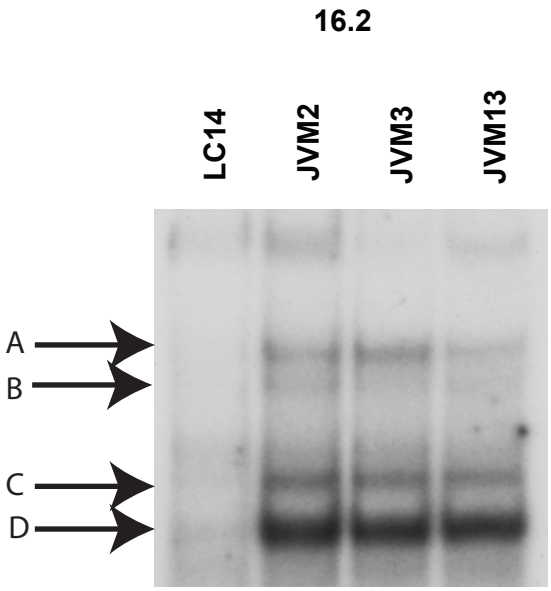


Figure 5.9A Autoradiograph of EMSA carried out with probe 16.2. The bands shifted on the EMSA are labeled A-D.

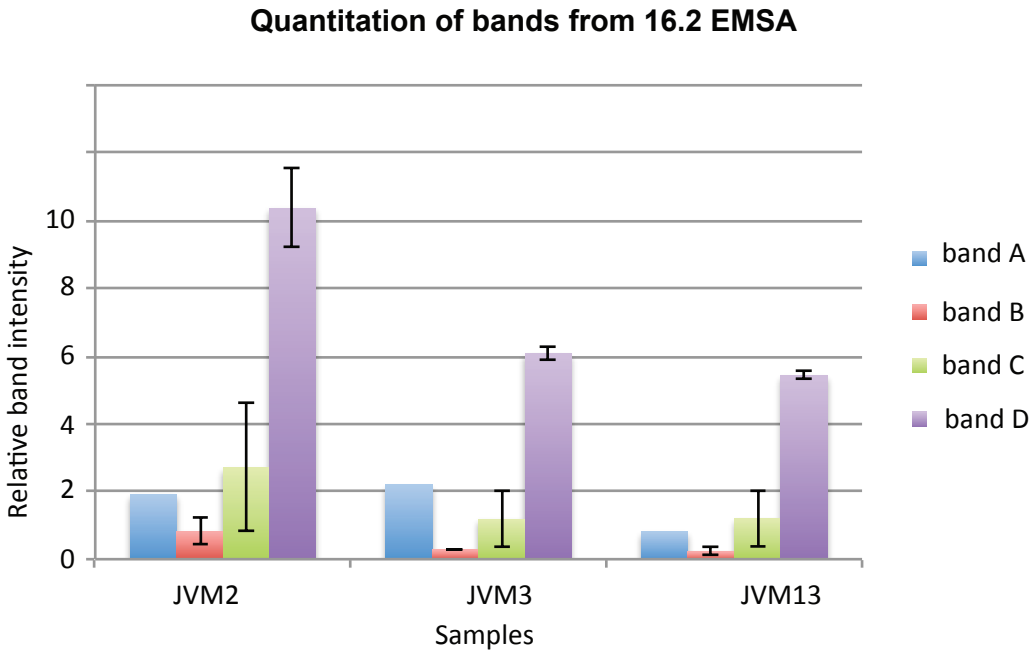


Figure 5.9B Quantitation from phosphorimager of the different bands. Each band is quantitated and then expressed relative to band B in Sp1, whose gel shift is shown on Fig 5.8A. The ratio of oligo band: Sp1 band for each CLL sample is then compared to the ratio of oligo band:Sp1 band for the normal sample, giving the relative band intensity. Error bars represent the SD from two separate gel shifts and quantitations. Band D corresponds to the most over-expressed protein, present at a level 6-10x greater in CLL cases than in normal CD19 cells.

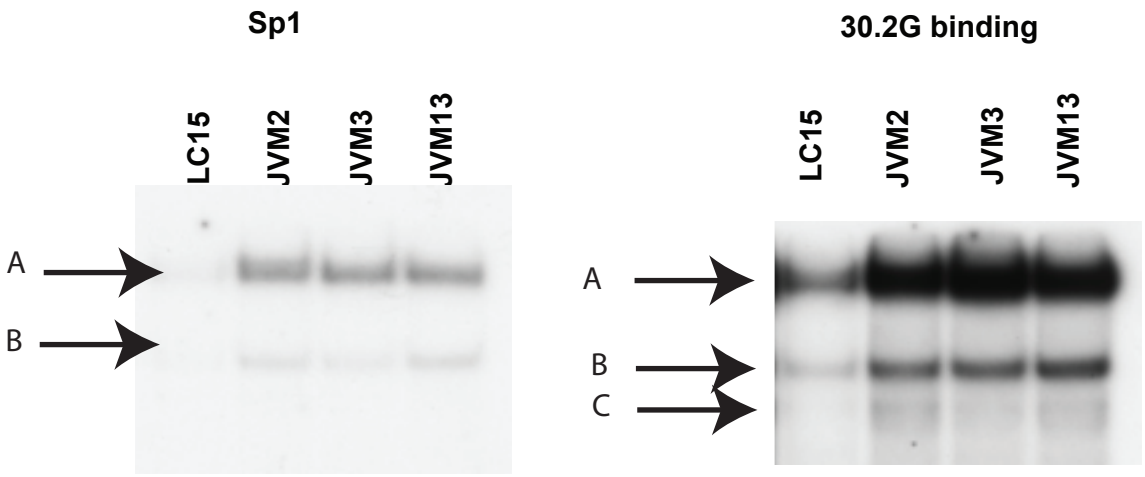


Figure 5.10A Autoradiograph of EMSAs carried out with probes Sp1 and 30.2. The bands shifted on the EMSA are labeled A-C.

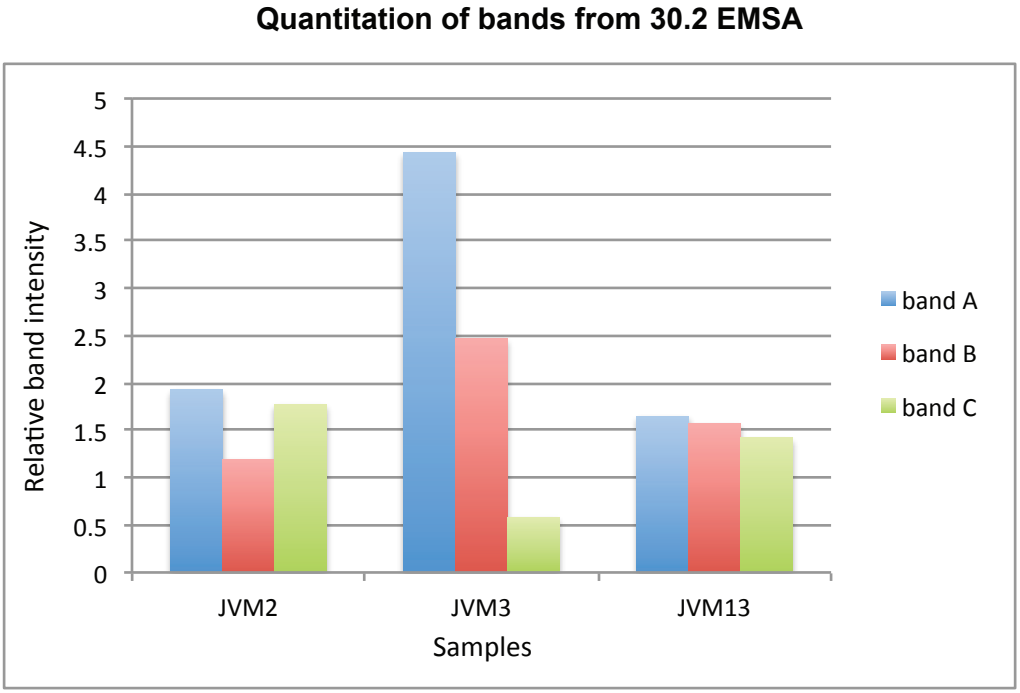


Figure 5.10B Quantitation from phosphorimager of the different bands. Each band is quantitated and then expressed relative to band B in Sp1. The ratio of oligo band: Sp1 band for each CLL sample is then compared to the ratio of oligo band:Sp1 band for the normal sample, giving the relative band intensity. Band A corresponds to the most over-expressed protein, present at a level 1.6-4.5x greater in CLL cases than in normal CD19 cells.

5.3.3 Primary CLLs

Although the protein(s) causing a gel shift were over-expressed in the CLL cell lines it was important to see whether this was also the case in primary CLL samples. The EMSA and quantitation are shown in Figure 5.11 and once again, compared to Sp1, most bands are increased in the CLL samples compared to normal. Some of the samples, however, have a biggest increase in band D (CLL-B and CLL-E) while CLL 137 and 147 have a biggest increase in band B, suggesting that there is heterogeneity between the CLL samples.

5.3.4 Methylated oligo

As the initial screening gel shifts were done with a probe containing an unmethylated CpG, the experiment was repeated with the same 16.2 probe but where the CpG was methylated, as it would be *in vivo*. The intensity of band D relative to the Sp1 band was compared to that for the same samples with the unmethylated 16.2 probe. As seen on Figure 5.12, methylation did not affect binding to the probe for samples JVM2 and JVM3, and slightly increased binding for JVM13. This would be compatible with the hypothesis of passive hypomethylation. If methylation of the probe had prevented binding, it would be more difficult to reconcile with the hypothesis.

5.4 DNA affinity chromatography

From these data, it appeared that the CLL cell lines contained one or several proteins that could bind to a sequence of DNA that contains a CpG which becomes unmethylated in CLL, and that this protein is over-expressed in CLL cell lines and primary samples compared to normal CD19 cells. The protein is still able to bind the probe in its methylated form, as the sequence would be *in vivo*. The next step was to identify this/these protein(s). The approach chosen was to use DNA affinity chromatography followed by liquid chromatography mass spectrometry (MITTLER *et al.* 2009a).

As illustrated in Figure 5.13A, the approach was to use biotinylated oligos which are attached to streptavidin beads in a column, run through the nuclear extract,

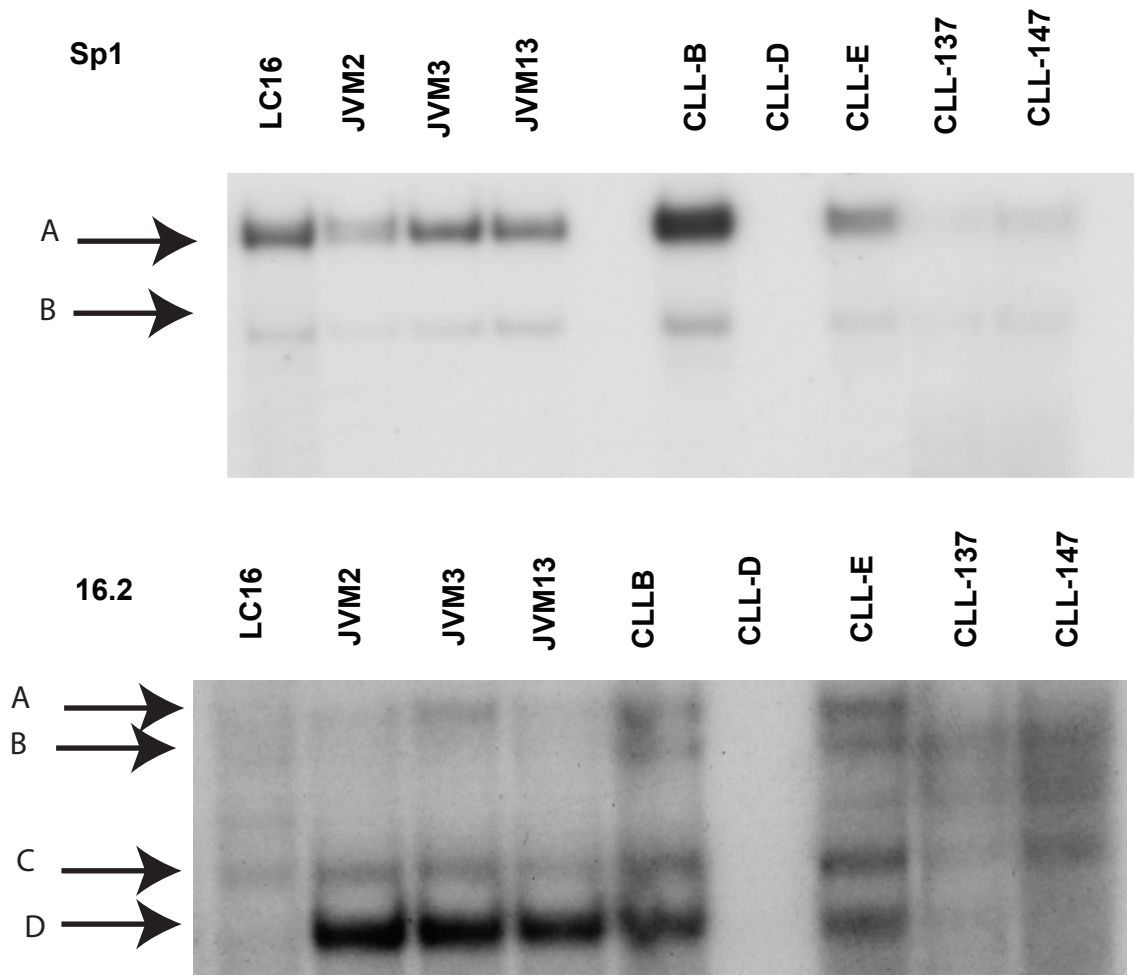


Figure 5.11A Autoradiographs of EMSAs carried out with two probes, Sp1 on the top and 16.2 on the bottom. Samples labeled LC are normal CD19 cells isolated from lymphocyte cones. JVM samples are CLL cell lines. The bands shifted on the EMSA are labeled A and B (Sp1) and A-D (16.2) as previously.

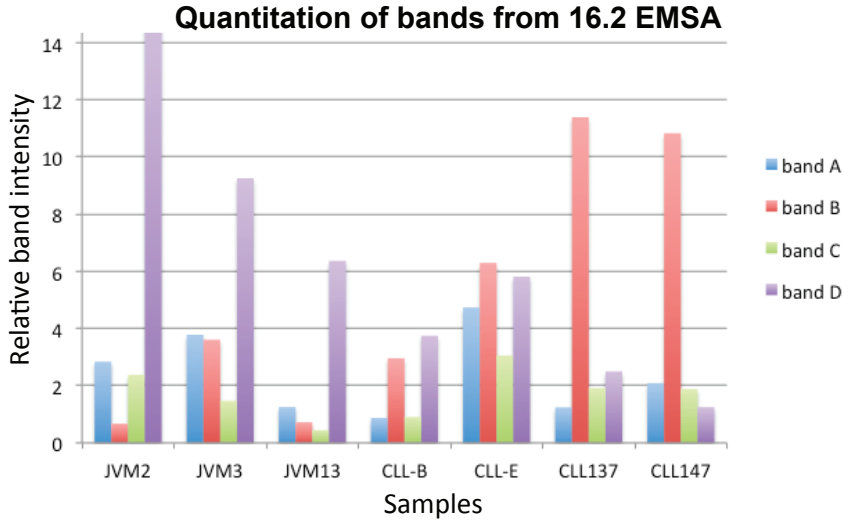


Figure 5.11B Quantitation from phosphorimager of the different bands. Each band is quantitated and then expressed relative to band B in Sp1. The ratio of oligo band: Sp1 band for each CLL sample is then compared to the ratio of oligo band:Sp1 band for the normal sample, giving the relative band intensity. While band D corresponds to the most over-expressed protein in the cell lines, the primary CLLs seem to have the greatest amount of the protein(s) causing the band B shift.

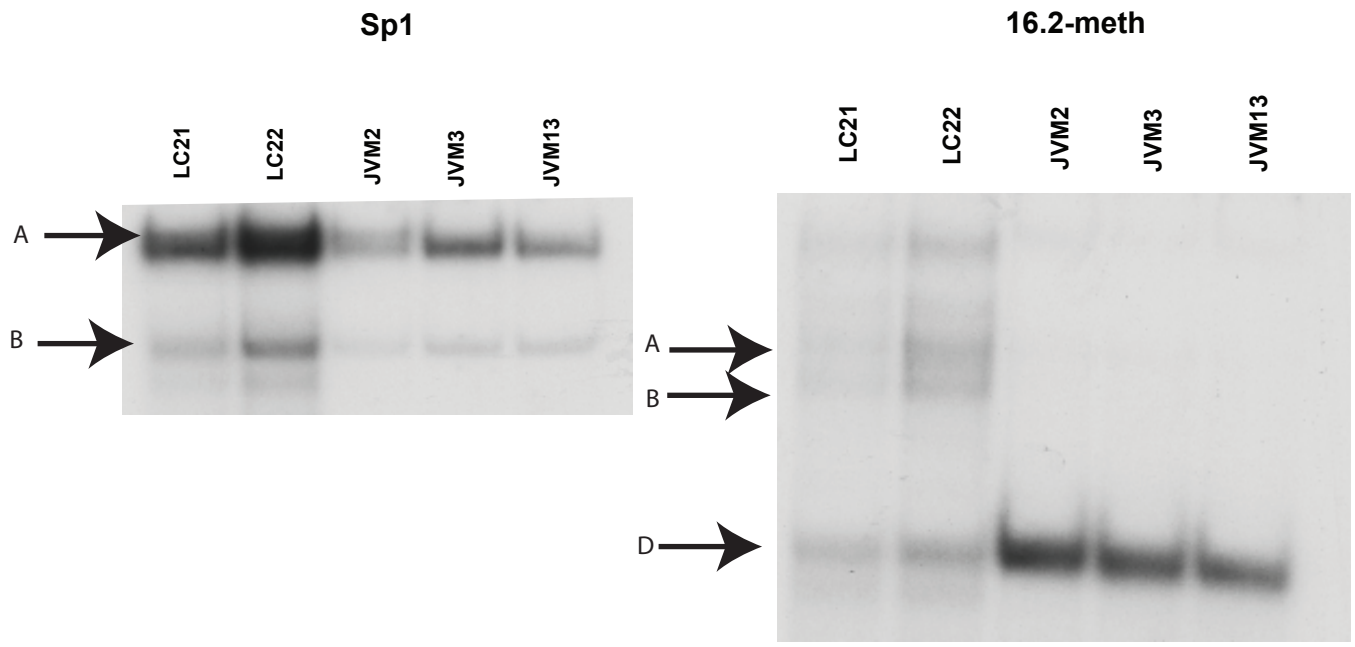


Figure 5.12A Autoradiographs of EMSAs carried out with two probes, Sp1 on the left and 16.2-meth on the right. Samples labeled LC are normal CD19 cells isolated from lymphocyte cones. JVM samples are CLL cell lines. The bands shifted on the EMSA are labeled A and B (Sp1) and A-D (16.2) as previously.

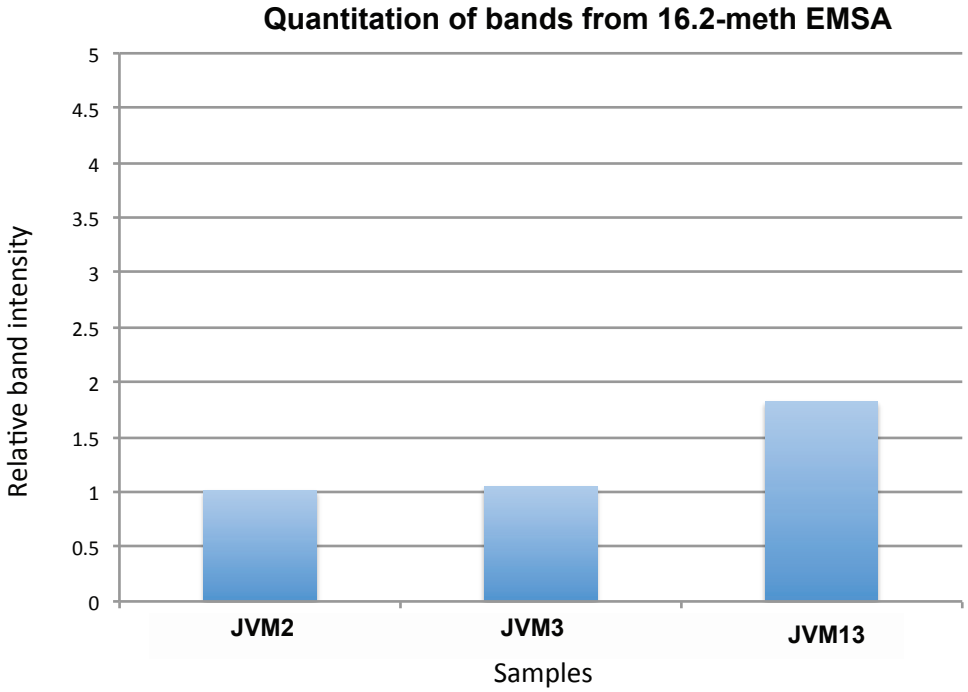


Figure 5.12B Quantitation from phosphorimager of band D. This is quantitated and then expressed relative to band B in Sp1, then the ratio of band D: Sp1 band for each CLL sample is then compared to the ratio of band D:Sp1 band for the same sample when 16.2 (unmethylated) is used, giving the relative band intensity. Methylation does not affect the binding in JVM2 and JVM3 but seems to slightly enhance the binding in JVM13.

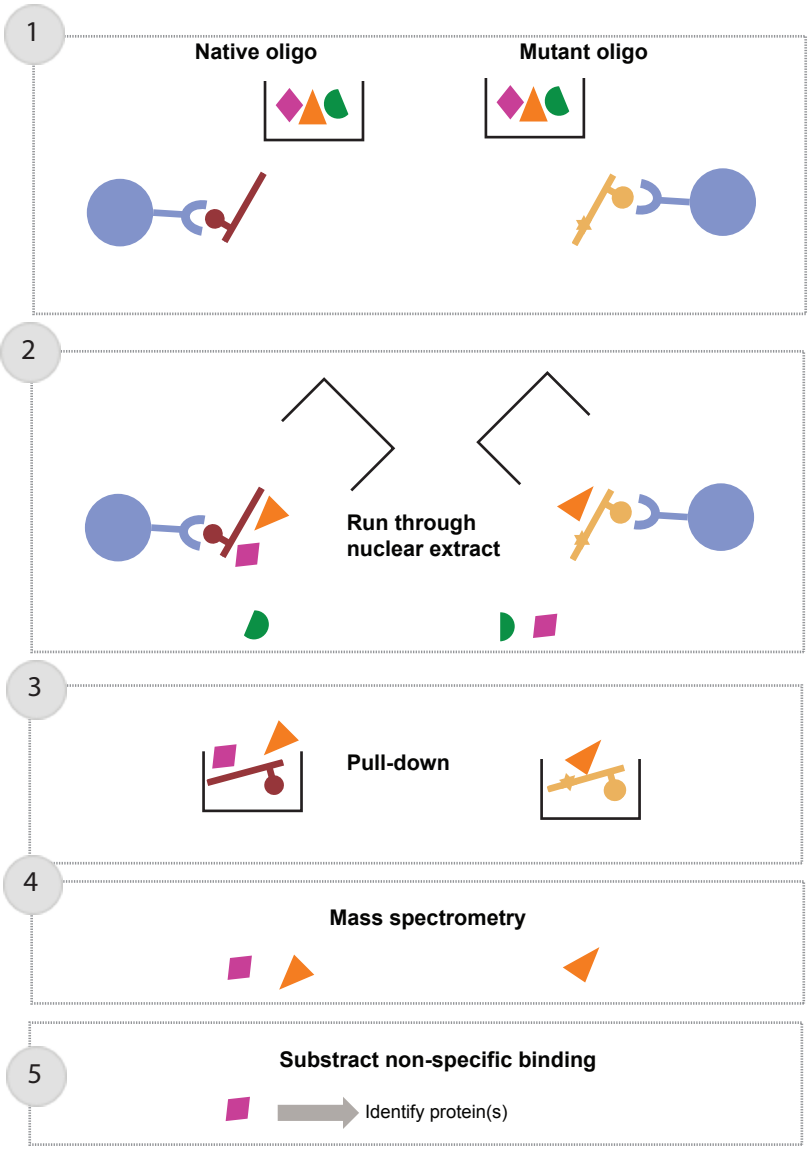


Figure 5.13A Schema of the steps involved in the DNA affinity chromatography method isolating the protein(s) responsible for the gel shift with oligo 16.2. (1) Native and mutant (abolished activity on EMSA) biotinylated oligos are attached to streptavidin-coated magnetic beads. (2) Nuclear extracts from CLL cell lines are run past both columns. (3) The proteins which have bound to the oligos in the columns are eluted. (4) Both fractions corresponding to the native and the mutant oligos are subjected to mass spectrometry (5) Proteins that are identified in both fractions are substracted from the analysis as they represent proteins which bind non-specifically to the oligo. Specific protein(s) are identified.

5'CAAAACTGCAGGAGCACCCACTAAACGGAAAAAAACAAACA3'

Figure 5.13B Forward strand sequence of the oligo used for pull-down. The red portion is the spacer sequence, the green portion is the Pst1 recognition site. The oligo is biotinylated at the 5' end of the forward strand.

Oligonucleotide name	Sequence
16.2	GGGAGCACCCACTAAA***CG***GAAAAAAACAAACA
16.2_mut1	GGGAG**A**ACCCACTAAA***CG***GAAAAAAACAAACA
16.2_mut2	GGGAGCAC**A**CACTAAA***CG***GAAAAAAACAAACA
16.2_mut3	GGGAGCACCCACT**G**AA***CG***GAAAAAAACAAACA
16.2_mut4	GGGAGCACCCACTAAA***TG***GAAAAAAACAAACA
16.2_mut5	GGGAGCACCCACTAAA***CG***GAG**A**AAAAACAAACA
16.2_mut6	GGGAGCACCCACTAAA***CG***GAAA**T**AAACAAACA
16.2_mut7	GGGAGCACCCACTAAA***CG***GAAAAAA**G**CAAACA
16.2_mut8	GGGAGCACCCACTAAA***CG***GAAAAAAACA**G**ACA

Figure 5.13C Table of the sequences for the mutant oligonucleotides. Mutated bases are highlighted in fuschia.

retain the oligos and their bound proteins, and then elute the proteins. After running these out on an SDS gel and in-gel tryptic digestion, the fragments were subjected to LC-MS and the identified peptides matched to known protein databases.

In this protocol, the elution step is done by including a PstI cutting site in the sequence. This results in the oligo still being attached to the protein(s) after the elution step. The sequence of the oligo used for pull-down is in Figure 5.13B.

5.4.1 Mutant oligos

In vitro binding assays using short DNA fragments are associated with high levels of non-specific protein binding. Identification of the specific protein responsible for the gel-shift by mass spectrometry would be hampered by the proteins that have bound non-specifically in the pull-down assay. Therefore, the protocol also involves running a parallel reaction with a similar oligo containing a mutation that diminishes binding of the specific protein. Such oligos can be chosen by identifying a reduction in the intensity of the gel shift band under investigation. As this oligo should retain much of the non-specific binding, when proteins associated with this oligo are analysed by mass spectrometry, they can then be subtracted from the analysis of the un-mutated oligo, increasing the likelihood of identifying the sequence-specific binding proteins. Eight mutants were therefore generated, with every 3rd base mutated from the original oligo. These are illustrated in Figure 5.13C, and a gel shift was carried out with all of these mutants used as probes as well as the native probe. As can be seen on Figure 5.14, mutants 1 and 6 had the greatest effect on band D, reducing binding to <20% of that with the native sequence. To ensure that these do not bind the same protein as that which binds 16.2 (as this would mean subtracting it from the analysis), it was checked whether mut1 and mut6 could displace 16.2 in a cold competitor assay. Unfortunately, as shown on Figure 5.15, they do displace it as much as a cold 16.2 does. This means that these would not be good controls, and it may be necessary to undertake the mutagenesis in 2bp/oligo instead of one. However, in the meantime progress was made to optimise the DNA affinity chromatography protocol.

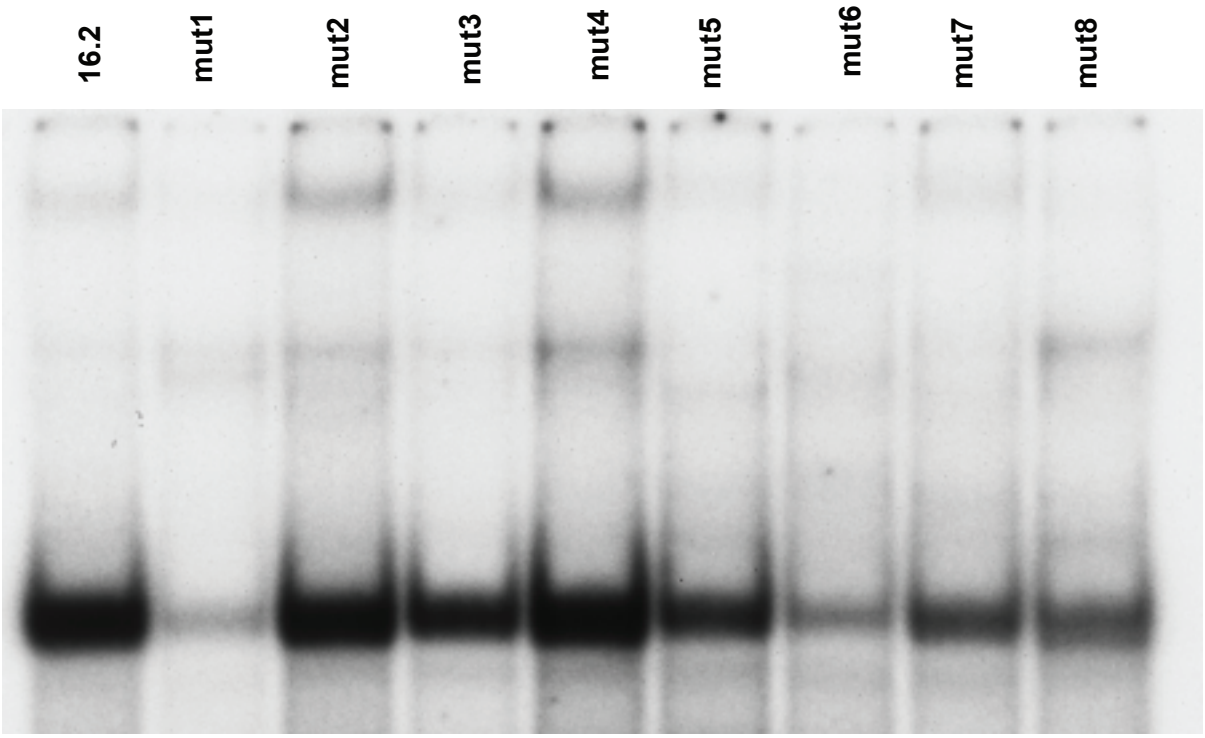


Figure 5.14A EMSA using a nuclear extract from a CLL cell line and the native 16.2 probe as well as mutants 1-8. From this gel, mutant 1 and mutant 6 appear to have the greatest reduction in binding compared to the native oligo.

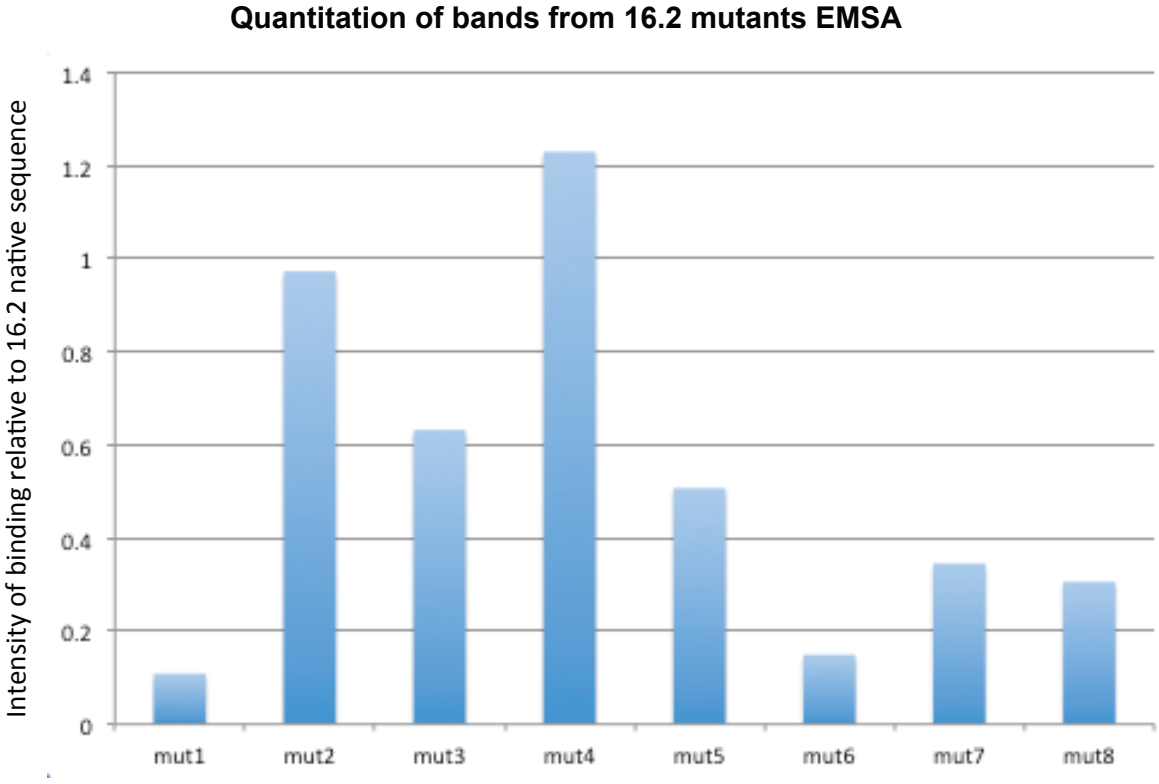


Figure 5.14B Quantitation of the above EMSA, confirming that the mutants with most reduced binding are mutants 1 and 6. The level of binding is quantitated relative to the band using the native 16.2 oligo as a probe.

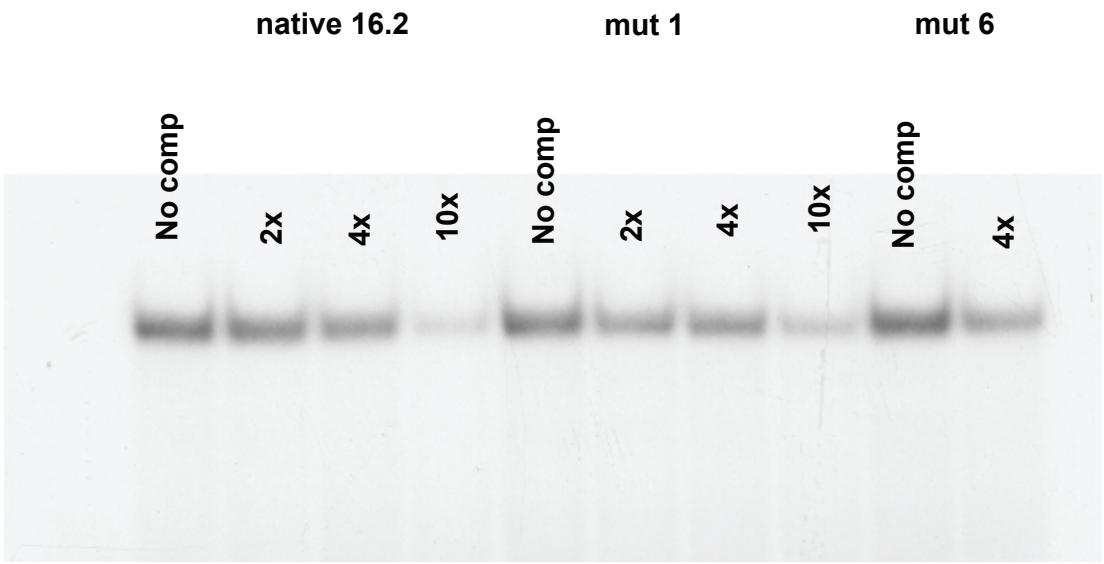


Figure 5.15A Cold competitor assay. The nuclear extract used was a CLL cell line and the radioactive probe was the native 16.2. The first 4 lanes show the results when the native probe was pre-incubated with the extract prior to incubation with the radioactive probe. The next 4 lanes show the results when mut1 was used as a cold competitor prior to incubation with the radioactive native 16.2 and the last two lanes for mut6. It can be seen that both mutants do compete with 16.2, meaning that the specific activity is also affected.

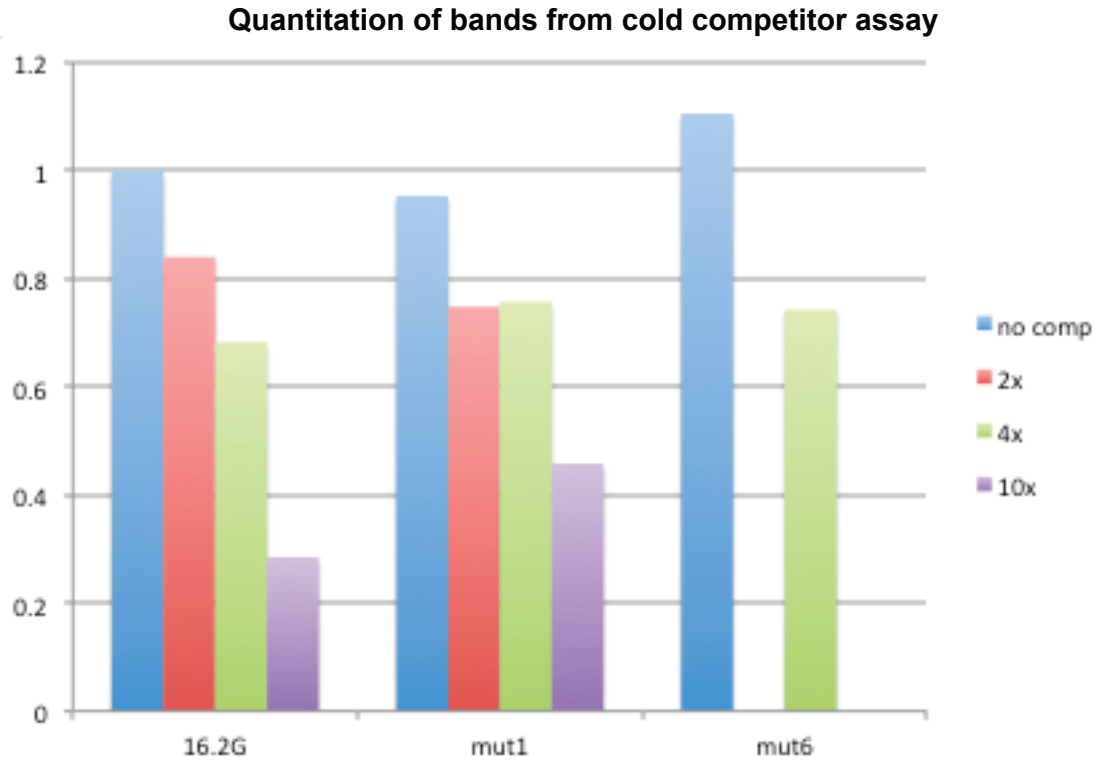


Figure 5.15B Quantitation of the above EMSA. This confirms that the mutants do displace 16.2 and that therefore they would not be good candidates to use in the pull-down assays as controls, because they would bind some of the specific protein and would therefore remove it from the analysis.

5.4.2 Optimisation of the DNA affinity chromatography protocol

5.4.2.1 Outline of the protocol

As outlined in Figure 5.16A, the protocol is as follows:

1. The oligo to be tested is designed with a biotinylated plus strand which contains a spacer sequence (5'CAAAA3'), a PstI recognition site (5'CTGCAG3') and then the oligo sequence. Complementary strands are annealed, with slight excess of the biotinylated strand, and allowed to bind to streptavidin beads. This is now the DNA affinity column.
2. The column is incubated with blocking buffer and washed.
3. The nuclear extract is adjusted to 10mM potassium glutamate and diluted with buffer G, which contains poly dIdC, then pre-cleared with naked streptavidin beads.
4. Once the naked beads are removed, the pre-cleared extract is mixed with the blocked column and incubated for 3 hours.
5. The beads (now containing the bound proteins) are isolated and the oligos detached from the beads using PstI digestion. This process is repeated once. Note at this point the eluted proteins are still bound to the oligo.

5.4.2.2 Plan to test the protocol

In order to test the protocol, different fractions were tested by EMSA. The fractions tested (as shown on Figure 5.16A) are:

1. "PRE": this is the nuclear extract to which only potassium glutamate and buffer G have been added.
2. "CLEARED": this is a sample of the supernatant that has been through pre-clearing. This supernatant is what will be exposed to the column.
3. "UNBOUND": this is the supernatant after incubation with the column. This should have very little of the protein being enriched and is the fraction that is discarded, as the protein should now be bound to the beads.
4. "BOUND1" and "BOUND2" are the results of the elution process after PstI cleavage.

Figure 5.16A Schema of the DNA affinity chromatography protocol. The nuclear extract is pre-cleared with naked beads and then incubated with streptavidin beads on which are immobilised biotinylated oligos (the column). After washing, the bound proteins are eluted with a PstI digest as the oligos contain a PstI recognition sequence.

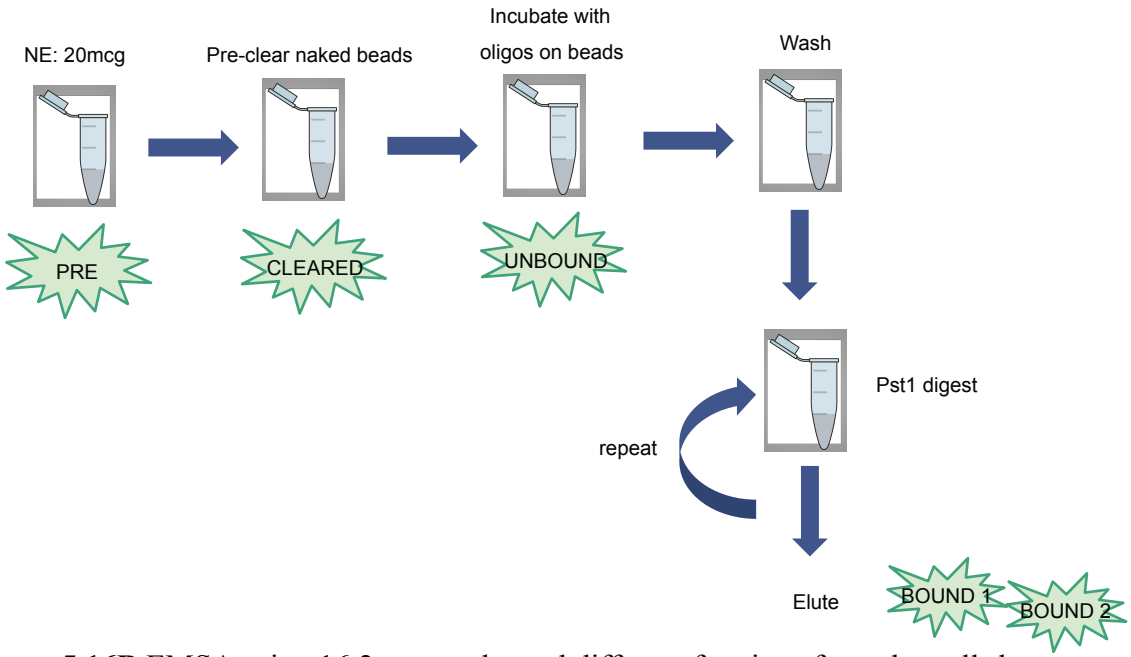
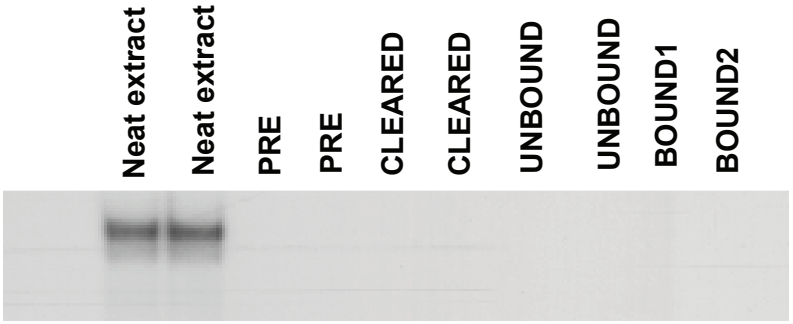


Figure 5.16B EMSA using 16.2 as a probe and different fractions from the pull-down. Only the neat extract works by EMSA. By the time potassium glutamate and buffer G have been added (PRE), the extract can no longer bind the probe.



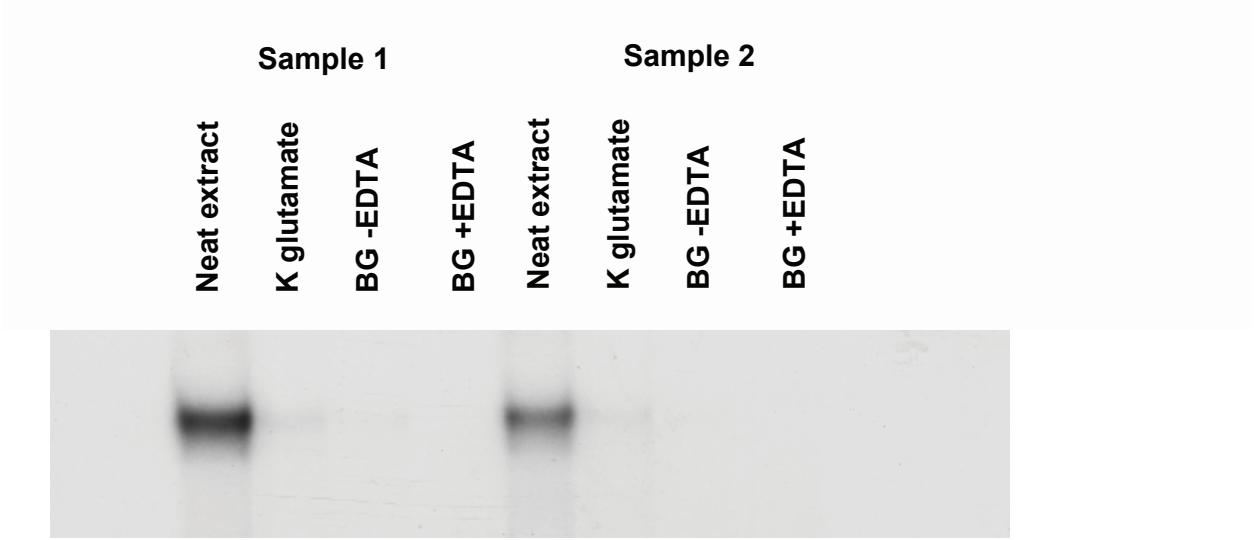


Figure 5.16C EMSA using 16.2 as a probe and neat extract incubated with varying buffers used in the pull-down. Potassium glutamate and buffer G both severely affect the binding. In buffer G it is not the EDTA which is reponsible for the majority of this effect, as removing EDTA did not dramatically improve binding.

It would be expected that PRE and CLEARED fractions would contain the protein and therefore to cause a gel shift in an EMSA with 16.2, whereas the UNBOUND should not. Finally, both BOUND fractions should cause a band shift by EMSA, as they should be enriched for the specific protein. As well as EMSA, the different fractions were tested on Coomassie or silver stains, where it would be expected that a ladder of decreasing complexity would appear as the samples go through the protocol.

5.4.2.3 Initial pull-down

The initial attempt at pull-down was not very successful, as the only fraction which worked on EMSA was the one with the nuclear extract alone (Fig 5.16B). The "PRE" fraction, containing potassium glutamate and buffer G is no longer able to bind the oligo. Potassium glutamate is used to decrease background binding, but at this concentration it may be inhibiting the binding of the protein to the oligo. Therefore the nuclear extract was tested before and after addition of potassium glutamate as well as with and without buffer G. As buffer G contains EDTA, which could be chelating zinc that could be necessary for DNA binding if the protein were a zinc finger protein, the nuclear extract was also tested with the buffer without EDTA. Figure 5.16C shows that potassium glutamate and buffer G (with and without EDTA) both affect the binding dramatically. The EDTA in buffer G cannot be responsible for the majority of this effect, however, as removing it from buffer G did not greatly improve binding.

5.4.2.4 Varying buffers

As buffer G also contains poly dIdC, different concentrations of this were tested- at 100% of the level decribed in the protocol, 10% and 1%. As seen on Figure 5.17, both 100% and 10% of the poly dIdC reduced binding (particularly of the upper bands) whereas 1% did not. However 10% poly dIdC still had a strong band D so this concentration was ultimately used so as to preserve some of its function in preventing non-specific binding rather than reduce it to 1% and lose most of its function. As can be seen in Figure 5.17, varying concentrations of NP40 were also tested as it was suggested that this is a low concentration and increasing it may remove some of the non-specific binding. Using up to 2% NP40 did not reduce the binding to the probe.

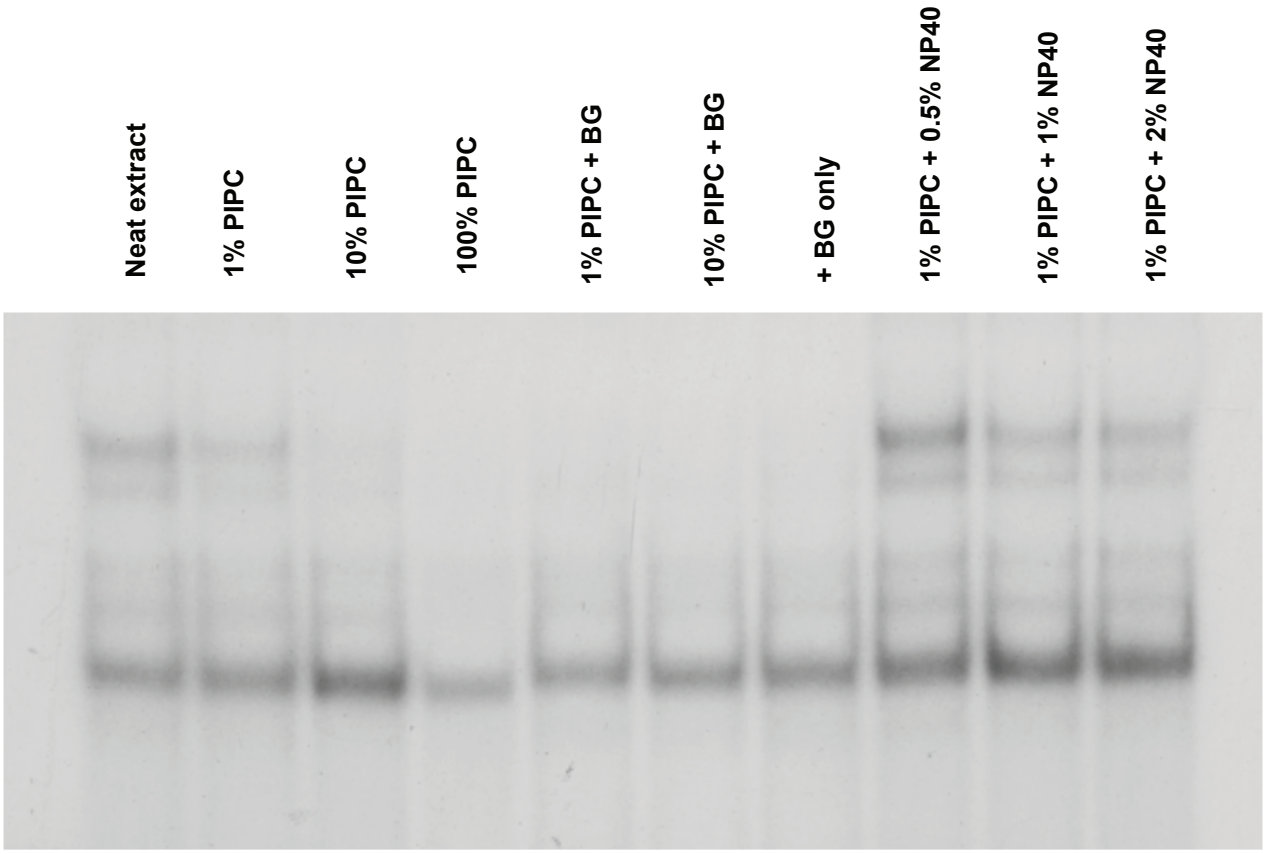


Figure 5.17 EMSA using 16.2 as a probe and neat extract incubated with varying concentrations of buffer constituents. Poly dIdC affects binding to the probe, with the greatest effect at 100% of that used in the protocol. An effect is still noted at 10% and no effect is noted at 1%. Once the concentration of Poly dIdC is reduced, the addition of the remainder of buffer G no longer abolished binding. It is possible to increase the concentration of NP40 4-fold without affecting binding.

5.4.3 Further optimisation

5.4.3.1 Second pull-down with altered buffers

The experiment was therefore repeated, removing the initial potassium glutamate, reducing poly dIdC to 10% and using two different amounts of NP40- 0.5 and 2%. The results are seen on Figure 5.18. It seems that these conditions are more compatible with binding in an EMSA assay, as although the addition of buffer G reduces binding to about 70% (and to about 30% with the higher concentration of NP40), there is some binding left after pre-clearance. Interestingly, the unbound fraction seems to have just as much binding as the pre-cleared fraction. The eluted fractions, however, show minimal binding in the EMSA assay. This could be due to a range of possibilities:

1. There is no binding of the protein in the nuclear extract to the bead-bound oligo. This could be due to a change in the conformation of the binding site once the oligo is attached to the linker and the bead. Alternatively the binding site may be too close to the bead, creating steric inhibition for the protein to bind the oligo.
2. The protein does bind to the bead and oligo but does not elute. This could occur if, once the protein is bound, it obscures the PstI site and the digestion reaction is unable to proceed.
3. The protein does bind the oligo and does elute from the bead, but at too low a concentration for it to work in EMSA.
4. The protein does bind the oligo and does elute at a good concentration but a constituent of the wash buffer prevents it from working in EMSA.
5. The presence of the cleaved oligo, which is still bound to the protein, prevents the protein from binding the radioactively labelled probe.

5.4.3.2 Determination of possible causes

To try to answer some of these questions, EMSA was carried out using the following conditions:

1. Serial dilutions of the nuclear extract, to simulate the protein being eluted at too low a concentration.

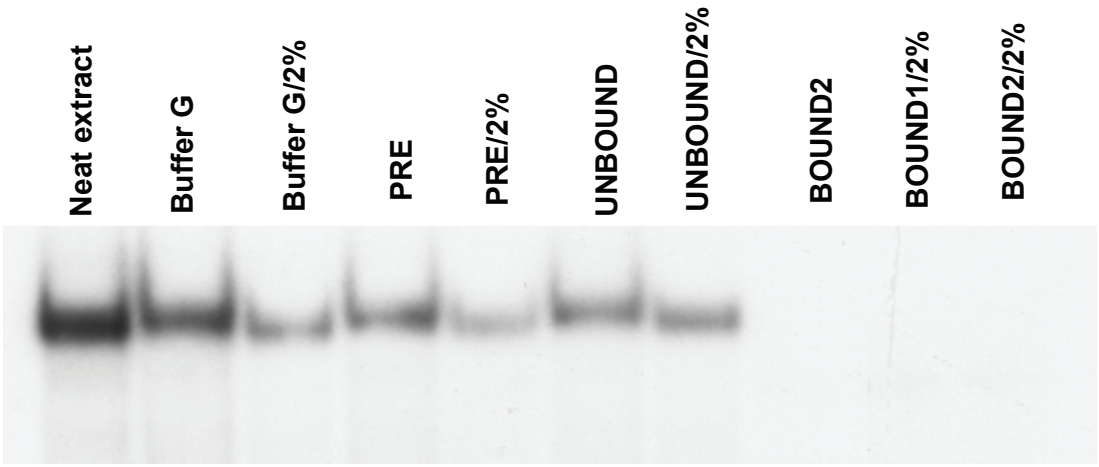


Figure 5.18A EMSA using 16.2 as a probe and different fractions from the pull-down and the altered buffer conditions as tested earlier. Now all of the fractions work by EMSA, with varying intensities.

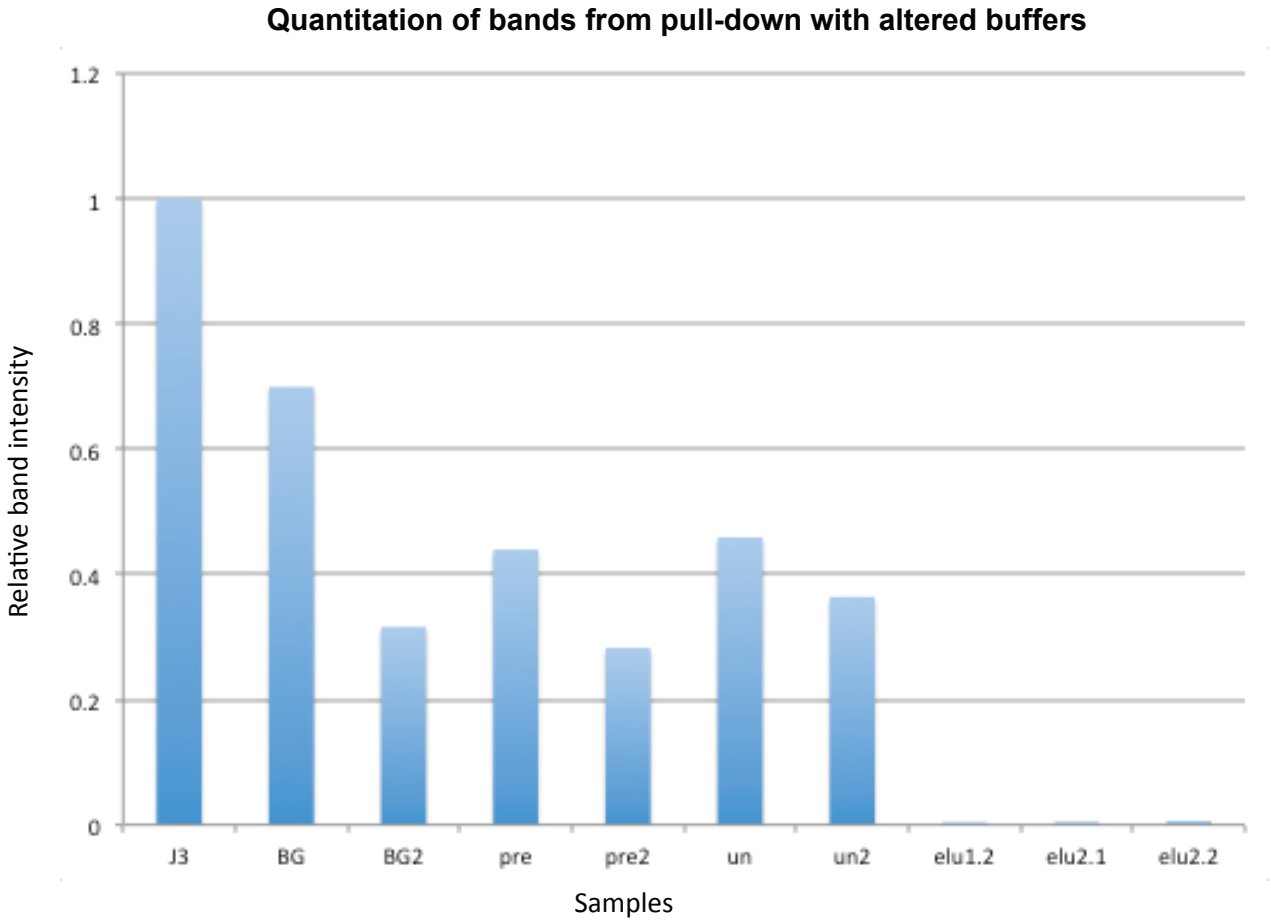


Figure 5.18B Quantitation of the above EMSA with the intensity of the bands relative to that in J3 (neat extract). Buffer G still reduces binding but to a lesser level. Whereas previously 2% NP40 did not appear to reduce binding, here it does. The amount of binding in the unbound fraction is as much as in the fractions prior to being subjected to the column. However the amount of binding from the eluted fractions is minimal.

2. Incubating the nuclear extract with 175pmol of oligo (biotinylated and not) prior to carrying out the EMSA. This would simulate the pull-down conditions as it should correspond to the amount of oligo that would be bound to the proteins following PstI cleavage.
3. Incubating the nuclear extract with the wash buffer prior to carrying out EMSA.

The results are shown in Figure 5.19A. Dilution of the extract 50-fold results in too low a concentration of the protein for binding in EMSA. Pre-incubation with the wash buffer does not prevent binding of the protein to the 16.2 probe. However, pre-incubation of the extract with 175pmol of oligo, which is the concentration at which the probe would be expected to elute with the protein following PstI cleavage, does abolish the reaction. This strongly raises the possibility that this is what is preventing binding by EMSA and that therefore EMSA may not be an accurate way of testing the eluted fraction for decision to carry through with protein identification from pull-down. It remains possible that the protein is eluted at too low a concentration to work by EMSA. Determining the actual concentration of protein in the eluted fractions by Qubit is, however, unhelpful, due to the presence of BSA and PstI in the eluted fraction. However, visualising the proteins on a silver stained gel would give an idea of whether or not there are adequate levels of proteins in the eluted fraction.

5.4.4 Silver stain

This is shown on Figure 5.19B and shows very nicely that there is protein in the eluted fractions and that, as expected, the complexity of the ladder is reduced. There also seem to be three prominent bands- A, B and C. B is only present in one of the eluted samples. Mass spectrometry was therefore used to examine whether these bands contained interesting proteins. This was done with gel extraction of each of the bands, in-gel trypsinisation and MALDI-TOF (Matrix Assisted Laser Desorption/Ionisation- Time of Flight) as detailed in Chapter 2.

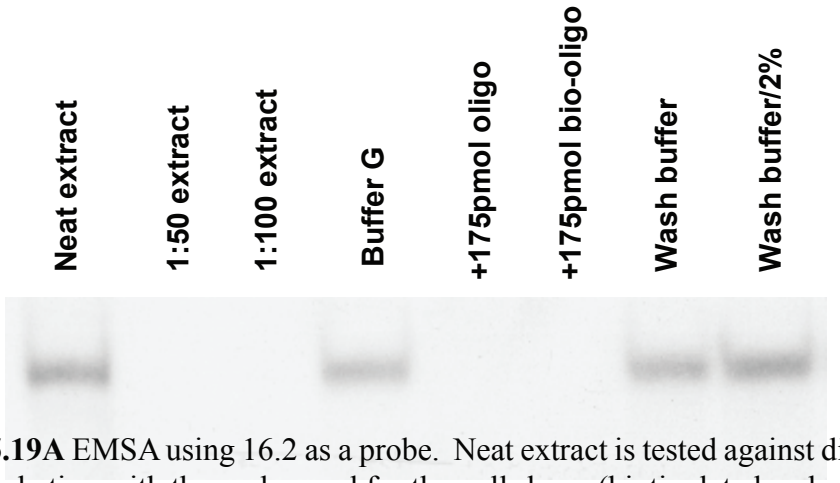


Figure 5.19A EMSA using 16.2 as a probe. Neat extract is tested against dilutions and against a pre-incubation with the probe used for the pull-down (biotinylated and not). Neat extract is also tested against a pre-incubation with the wash buffer. 1:50 and 1:100 dilutions of the extract are too low concentrations for binding to be picked up by EMSA. Both the biotinylated probe and the unbiotinylated one completely abolish binding to the 16.2 probe, suggesting this may be the explanation for lack of a band in the eluted fractions following pull-down. The wash buffer (with two separate concentrations of NP40) does not affect binding in EMSA.

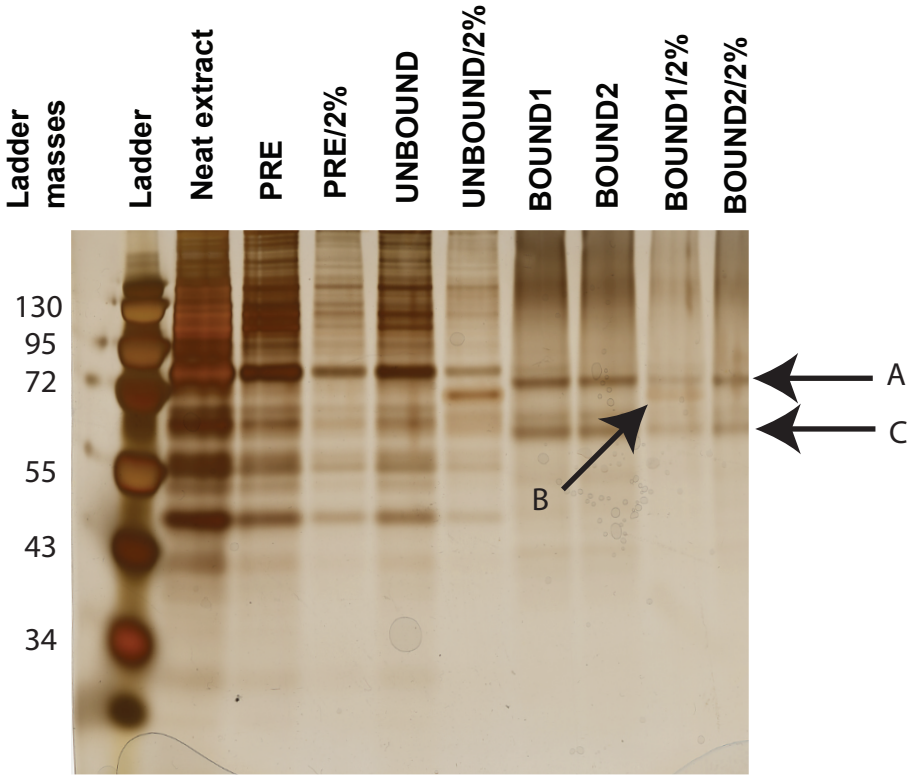


Figure 5.19B Silver stain of the fractions in the pull-down. The bound fractions are of greatly reduced complexity, suggesting numerous proteins have been selected out by the process. Three bands appear strongly in the bound fractions, labeled A-C, although a number of other bands are also visible.

5.4.5 MALDI-TOF results

The three bands were analysed separately. From the MALDI-TOF analysis, a series of peptide ion masses were logged for each band. Mascot then performs peptide mass fingerprinting, by using a database of peptide ion masses from a protein digest against the UniProt database in order to predict the sequence of the protein. The Mascot report gives a protein score, which corresponds to the probability of the match being random. Pure proteins will be identified at high probability ($p<0.05$) however a protein could still be the correct match and yet not fall in the $p<0.05$ probability if the purity is suboptimal. It is, however, difficult to know which is the correct protein if none of the hits is statistically significant. From a Mascot histogram, as shown in Figure 5.20, the proteins with a score above the threshold value of statistical significance fall outside the green shaded area, whereas those below the statistical significance threshold lie in the green shaded area. Figure 5.20A shows that band C was unequivocally identified as keratin, which must be a contamination in the sample (score 73, threshold being 70). Similarly, band A was identified as BSA, which is present in the bead blocking buffer.

Band B, however, which is present in only the first eluate of the reaction carried out at a higher NP40 concentration (and which should therefore have less non-specific binding) does not contain an unequivocal protein. As can be seen in Figure 5.20B, all the potential hits fall in the green shaded area and are outside the statistical significance threshold. Interestingly, band B is only present in the first eluate, where the specific protein should be at a higher concentration than in the second eluate. Looking through the non-statistically significant matches, several candidate proteins emerge. The most interesting candidate is MBD4 and the Mascot report is illustrated in Figure 5.20C. This has a predicted mass of 66.8kDa. Six peptides matched MBD4 exactly, covering 12% of the protein. Another candidate is hnRNP (heterogeneous nuclear ribonucleoprotein). Although this is normally an RNA-binding protein, involved in splicing, it has been reported to bind to DNA (FACKELMAYER *et al.* 1994) and if it were the over-expressed protein in question could well bind aberrantly to DNA whereas it would normally be an RNA-binding protein. The protein from band B matched 8

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 70 are significant (p<0.05).

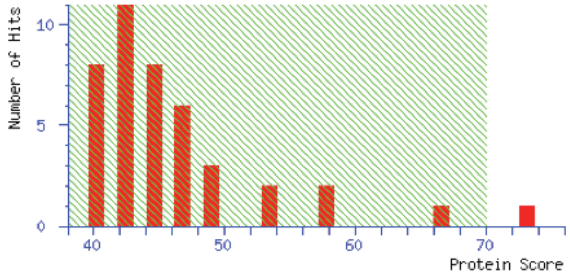


Figure 5.20A Mascot return for band C. Note that one protein has a score in the statistically significant range (outside the green shaded area). This protein was unequivocally identified as keratin.

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 56 are significant (p<0.05).

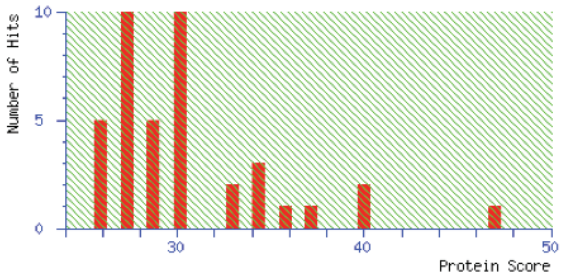


Figure 5.20B Mascot return for band B. Note that no proteins are unequivocally identified here as there are none with a probability score that falls in the statistically significant range and all possibilities are in the green shaded area.

Taxonomy: Homo sapiens

Fixed modifications: Carbamidomethyl (C)
Variable modifications: Oxidation (M)
Cleavage by Trypsin: cuts C-term side of KR unless next residue is P
Number of mass values searched: **28**
Number of mass values matched: **6**
Sequence Coverage: **12%**

Matched peptides shown in **Bold Red**

```
    1 MGTTGLESLS LGDRGAAPTV TSSERLVPDP PNDLRKEDVA MELERVGEDE
   51 EQMMIKRSSE CNPLLQEPIA SAQFGATAGT ECRKSVPCGW ERVVKQRLFG
  101 KTAGRFDVYF ISPQGLKFRS KSSLANYLHK NGETSLKPED FDFTVLSKRG
  151 IKSRYKDCSM AALTSHLQNQ SNNSNWNLRT RSKCKKDVFM PPSSSSELQE
  201 SRGLSNFTST HLLLKEDEGV DDVNFRKVRK PKGKVTILKG IPIKKTKKGC
  251 RKSCSGFVQS DSKRESVCNK ADAESEPVAQ KSQLDRTVCI SDAGACGETL
  301 SVTSEENSLV KKKERSLSSG SNFCSEQKTS GIINKFCSAK DSEHNEKYED
  351 TFLESEEIGT KVEVVERKEH LHTDILKRGS EMDNNCSPTR KDFTGEKIFQ
  401 EDTIPRTQIE RRKTSLYFSS KYNKEALSPP RRKAFKKWTP PRSPFNLVQE
  451 TLFHDPWKLL IATIFLNRTS GKMAIPVLWK FLEKYPSAEV ARTADWRDVS
  501 ELLKPLGLYD LRAKTIVKFS DEYLTKQWKY PIELHGIGKY GNDSYRIFCV
  551 NEWKQVHPED HKLNKYHDWL WENHEKLSLS
```

Figure 5.20C Mascot match to MBD4. This match is not statistically significant below 0.05, however no protein was identified with p<0.05 for that band. Six peptides match exactly, covering 12% of the protein.

peptides to hnRNP, covering 18% of the sequence. It has a predicted mass of 69.8kDa.

However, it must also be noted that one of the proteins which had equal likelihood of matching is again keratin (65.6kDa), with 7 matching peptides (covering 11% of the protein). Therefore, although this is not like band C, where keratin was unequivocally identified, it remains possible that band B is neither MBD4, a tantalizing candidate protein for being involved in DNA demethylation, nor hnRNP, which it could be imagined might be able to bind to DNA under malignant conditions, but is merely another keratin contamination.

5.5 Discussion and further work

At this stage, it is not clear whether the proteins identified by MALDI-TOF are genuine candidate proteins for the bands seen on EMSA using probe 16.2. Furthermore, even if these are genuine proteins present in the nuclear extract and which have bound the oligo, it does not follow that they are the proteins responsible for the bands in the 16.2 EMSAs, as they could represent non-specific binding to the oligo.

Several avenues of investigation are now open, which will be detailed below.

5.5.1 Pursuing the silver stain and MALDI-TOF avenue

Because of the tantalizing results of the MALDI-TOF from the silver stain band B, the first step would be to repeat the pull-down reaction and see whether band B appears again, and whether MALDI-TOF results are similar. If so, it would be worth investigating whether MBD4 in particular is the protein in question. This could be done by using an antibody against MBD4 in a supershift assay using CLL nuclear extract and probe 16.2. If MBD4 is the protein responsible for the band shift, the band will be further retarded with addition of the antibody. If, on repeating the experiment, band B is now unequivocally identified as keratin (or does not appear in the repeat experiment), other much fainter bands may be worth following-up and it may be that the three strong bands are all either

contaminants or proteins added in the buffer. Ideally, keratin contamination could be prevented in a repeat of the experiment. Removing BSA would be more difficult. It would be possible to test different concentrations of BSA in the buffers or perhaps try leaving it out altogether. Alternatively, it would be possible to investigate a method for removing BSA from the sample after the elution. Another possibility for improving the identification of bands other than A, B and C from the eluates would be to increase their concentration by repeating the experiment several times and pooling the samples together.

5.5.2 Trying to pursue the use of EMSAs as a monitoring assay

Because it is possible that the protein is present in the eluate but that it simply does not work by EMSA, it would be useful to determine whether there is a way of getting the EMSA to work. If the problem is with the oligo still being present on the protein and preventing binding to the radioactive probe, it may be possible to elute the protein without doing the PstI digestion. It would be worth investigating whether a method such as altering the pH could dislodge the protein from the DNA column.

Another possibility would be to repeat the entire experiment using the oligo to Sp1 in the column. This would mean that once the protein has been eluted with PstI cleavage, it would be possible to test the eluate for the presence of Sp1 using Western blotting. If Sp1 is seen by Western blotting but the EMSA with an Sp1 probe does not work, this would confirm that EMSA cannot be used as a test for the protein prior to LC-MS. It is, however, possible that doing this test with Sp1 as a probe would not entirely replicate the problems occurring with oligo 16.2 as a probe. For example, if the problem with 16.2 is that there is a conformation change induced by the attachment of the oligo to the linker and the bead and the protein can no longer bind its recognition site, or if once it is bound it obscures the PstI cleavage site, this does not mean that the same would happen with the Sp1 oligo and therefore the Sp1 EMSA would work and not reveal what the problem is with 16.2.

5.5.3 Optimising the original protocol

It may not be possible to work out whether EMSA is a good way of testing whether the protein is still in the eluate, and it is possible that on repeating the silver stain there would not be a particular band worth pursing by MALDI-TOF. However, the original pull-down protocol still remains to be optimised, in particular the identification of a mutant that abolishes the specific binding. To do this, double mutants for the oligo should be tested until one is identified that does not bind the protein that causes a band shift with 16.2 and does not compete for 16.2. At this point it would then be possible to carry out the pull-down assay in parallel with 16.2 and the mutant and subject both eluates for LC-MS with the argument that the non-specific binding proteins would be eliminated from the analysis. Not using EMSA as a test to confirm the presence of the protein that causes a band shift with 16.2 in the eluate obviously carries the risk that the LC-MS would not yield any result, or more likely that all candidates suggested by LC-MS would be spurious. However this may be the only avenue left to pursue if the others fail to yield a convincing alternative.

The pieces of evidence which argue in favour of pursuing this path are that a protein was identified as quantitatively over-expressed (5-10 fold) in CLL cell lines and primary patient samples compared to normal CD19 cells. This protein or protein complex binds an oligo corresponding to the 8th most hypomethylated CpG in CLL, and this CpG is consistently and statistically significantly hypomethylated in a population of CLL samples. The protein can bind the oligo when it is methylated, as we would expect it to be *in vivo*. While there is no direct evidence that the pull-down reaction is working, if there was a parallel reaction with a mutant oligonucleotide, it would be worth comparing these two separate fractions by LC-MS, as this may reveal the over-expressed protein in CLL able to bind the sequence corresponding to a hypomethylated CpG.

Chapter 6. Final discussion and conclusions

In attempting to answer a scientific question, the tools employed will shape, if not distort, the answer that can be reached. Initial tools that were used to study DNA methylation in malignancy established crucial patterns, such as hypomethylation at repetitive elements and hypermethylation at a number of CpG island promoters, which have since been expanded and refined.

In today's post-genomic era, 'genome-wide' studies of DNA methylation have proliferated, with varying levels of breadth and depth of coverage. Indeed, the techniques employed in this study have also covered a range of sensitivities and specificities. The detailed analysis presented here of a seemingly promising genome-wide methylation assessment technique, MeDIP, highlighted the importance of careful and thorough evaluation of a method before starting to draw any conclusions from data it can generate. The inability of MeDIP to identify CpG islands as less than 30% methylated, a failing akin to that of its inability to distinguish a fully methylated from a fully unmethylated cell line (IRIZARRY *et al.* 2008b) invalidates its usefulness as a tool with which to probe in detail the altered patterns of methylation in malignancy.

My involvement in the development of a targeted bisulfite sequencing (TBS) technique occurred in an 'interim' climate where whole genome bisulfite sequencing is still unaffordable for the sequence coverage needed, and targeted sequencing can provide the equivalent accuracy, albeit in a much smaller region. While this technique, which has been independently developed by a number of groups, may become obsolete when costs of whole genome sequencing allow it to gain wider access, its current use has allowed the identification of a number of processes requiring optimisation in anticipation for greater levels of high throughput sequencing data. For example, it has spurred the development of bioinformatic tools that can now adequately deal with mapping in the context of a reduced complexity genome. The validation of TBS carried out within the context of this thesis justifies its use as a tool in the analysis of DNA methylation to answer different questions currently under investigation in our lab.

The wide range of genome-wide approaches to analysing DNA methylation, despite their individual failings and biases, have all participated in clarifying some issues regarding DNA methylation anomalies in malignancy. Their wider coverage and their ability to link to patterns of histone modifications have shed light on the links between DNA methylation changes in malignancy and key events occurring during development (HANSEN *et al.* 2011; KESHET *et al.* 2006; SCHLESINGER *et al.* 2007; SPROUL *et al.* 2011; VIRE *et al.* 2006). While it now seems that the DNA methylation occurring in malignancy is either not involved in silencing the genes affected, or at least not the main mechanism for doing so (SPROUL *et al.* 2011), the non-random overlaps of t-DMRs (tissue-specific differentially methylated regions) and c-DMRs (cancer-specific differentially methylated regions) with the targeting of developmentally regulated genes for methylation in cancer (IRIZARRY *et al.* 2009) suggests one of two explanations. Either this merely reflects the cell type in which malignancy arises, such that genes that are differentially regulated in that lineage become differentially methylated in that malignancy. Alternatively, malignant cells may re-acquire epigenetic control mechanisms which allow them to develop or maintain self-renewal. A transcription factor hypothesis has been suggested as a mechanism for this phenomenon of coordinated hypermethylation. As some transcription factors can interact with DNMTs to target methylation to their binding sites (HERVOUET *et al.* 2010b), the coordinated hypermethylation observed in malignancy would stem from the expression pattern of different transcription factors, which would logically be lineage-specific.

Global hypomethylation has long been ascribed to hypomethylation at repetitive elements, the result of which is an environment that favours chromosomal recombinations and the accumulation of genetic lesions, providing some cells with a selective advantage. Proposed mechanisms for this finding include low levels of DNMT3 isoforms or lack of its normal recruitment to DNA through loss of interacting partners such as LSH (EHRLICH 2009). Recent findings that hypomethylation occurs in blocks in the genome and that these blocks correspond to regions of the genome that become hypomethylated upon

differentiation (HANSEN *et al.* 2011) once again suggests a link between the patterns of methylation seen in differentiation and those imposed in malignancy.

Both active and passive mechanisms of hypomethylation in malignancy have been suggested. Some authors, based on studies of hemi-methylated CpG dyads at repetitive elements and statistical modeling, favour a hypothesis of active demethylation in malignancy (EHRLICH 2009). It is probable, however, that a number of mechanisms will be uncovered to explain different aspects of methylation anomalies in malignancy. It seems likely that 'epimutations', just as somatic gene mutations, will arise from a variety of mechanisms.

The data presented in this thesis support the hypothesis that some epigenetic changes in malignancy may arise via passive hypomethylation. To summarise, particular CpGs in the CLL samples studied are targeted for hypomethylation, although there is some variability between individuals. The loss of methylation is partial (on a population basis) and occurs particularly at sites of DNA that are accessible in the context of chromatin. The association between sites of hypomethylation and certain DNA motifs suggests that these could play a role in directing such changes. Regions enriched for transcription factor binding sites are more prone to hypomethylation while those depleted for such sites are less frequently targeted. Hypomethylation does not seem to be directed (or substantially reflected) by the activating histone modifications studied. The hypothesis tested in this thesis is that, in CLL, aberrant binding to these DNA sequences by proteins that are overexpressed in malignancy obscure the CpGs involved from the maintenance methylation machinery. The initial hemi-methylation that results would be perpetuated at the next stage of cell division. To this end, I have shown that some proteins that can bind sequences that become hypomethylated in malignancy are overexpressed in CLL. Although in the timeframe of the project it was not possible to identify such a protein, progress has been made in this direction.

What external support can be found for this hypothesis? Certainly, it is not commonly discussed in the literature. For example, a recent review on genome-

wide hypomethylation in malignancy proposed several hypotheses for passive demethylation (WILD and FLANAGAN 2010), including only a low level of substrate for the methylation reaction (SAM), low levels of DNMT1, mis-targeting of DNMTs or DNA methylation loss secondary to primary anomalies in histone modifications. It is, however, well supported from the developmental field, where hypomethylation of CpGs in transcription factor binding sites have been described. For example, depletion of DNA methylation was noted at general and specific transcription factor interaction sites with DNA (LISTER *et al.* 2009). *In vitro* experiments also support a passive loss of DNA methylation at sites of DNA-protein binding (LIN and HSIEH 2001; LIN *et al.* 2000; LORIOT *et al.* 2008). More recently, a series of studies on the MAGEA1 gene (a cancer-germline gene normally methylated in somatic tissues and unmethylated and expressed in malignant tissues) have shown that hypomethylation in malignancy may be the "footprint" of a transient passive demethylation event. The replication-dependent demethylation is ascribed to a transcriptional regulator interfering with the maintenance methylation machinery (DE SMET and LORIOT 2010), a rare mention of the hypothesis I have indeed been testing in this project. Interestingly, the predilection of transcription factor binding sites for being affected by cancer-linked hypomethylation has recently been reported (HANSEN *et al.* 2011). Furthermore, repetitive sequences are frequently hypomethylated and known to harbour multiple transcription factor binding sites (POLAK and DOMANY 2006).

A fascinating study of the effect of transcription factor binding sites on DNA methylation has just been published (LIENERT *et al.* 2011). In a series of detailed and elegant experiments, the authors inserted a variety of promoter sequences into an inert locus in mouse ES cells, namely beta globin. This eliminated the effects of transcription and the chromosomal environment, as there were no activating or repressive histone marks at this locus. They found that short (~1kb) sequences within the promoters, could accurately recapitulate the endogenous methylation pattern of each of these promoters. For example, insertion of the *Nanog* promoter resulted in an unmethylated state in undifferentiated cells and a methylated state upon neural differentiation,

identical to that found at the endogenous *Nanog* locus. Furthermore, shorter sequences within these regions, corresponding to transcription factor binding sites, were sufficient to correctly dictate the DNA methylation patterns, an effect that was abrogated by mutation of the binding sites. This phenomenon occurred not only for methylated promoters which maintained methylation upon differentiation, but also methylated promoters which lost methylation, unmethylated promoters which stayed as such and unmethylated promoters which gained methylation. This suggests that there may after all be a link between hyper- and hypomethylation, if a mere sequence corresponding to a transcription factor binding site can direct the correct methylation pattern normally ascribed to this sequence. Interestingly, however, no promoters were tested which are methylated in ES cells and lose methylation upon differentiation, a situation reminiscent of that occurring in the context of malignancy. Nevertheless, the fact that short DNA sequences are necessary and sufficient to recapitulate DNA methylation states is very interesting. An understanding of which *cis* or *trans* acting factors translate the sequence information into the correct DNA methylation state will pave the way towards understanding how these transcription factors and their interacting partners are altered to produce the well-recognised DNA methylation anomalies seen in malignancy.

Finally, a surprising recent report suggests that as cells become terminally differentiated, in this case erythroid cells, they undergo massive hypomethylation (25-30% loss) on a global and localised scale (SHEARSTONE *et al.* 2011). This process was not dependent on DNMT1, DNMT3 or GADD45a but dependent on cell cycling and DNA replication, implicating passive demethylation. The authors suggest that this global loss of methylation in terminal differentiation may be important for the correct expression of massively expressed genes such as erythroid genes in this context. Indeed, inhibiting demethylation reduced β-globin expression. It is intriguing to consider whether this mechanism could be active in malignancy and whether the loss of methylation seen in the malignant context partly reflects the rapid cell

division and the induction of genes involved in this process (cell cycling, anti-apoptotic, etc).

In conclusion to the work presented in this thesis, after an initial period of assessing and optimizing different techniques for analysis of DNA methylation, I was able to gather evidence suggesting a passive demethylation process in chronic lymphocytic leukaemia, as well as making progress towards identifying proteins which could be involved in this process. As well as pursuing the avenues employed and which have not been exhausted, further work could also test the hypothesis directly by over-expressing a transcription factor in malignant cells and observing whether this results in progressive hypomethylation at its binding sites. Understanding the mechanism for hypomethylation in malignancy could yield insight into both the normal regulation of the DNA methylation machinery and key processes involved in the development of malignancy.

Appendix 1: Primers and antibodies

A. RQ-PCR expression primers. All from Invitrogen, Taqman chemistry

Specificity	Reference
DNMT1	Hs00154749_m1
DNMT3a	Hs01027166_m1
DNMT3b	Hs00171876_m1
DNMT3L	Hs01081364_m1
18S	Hs03928990_g1
GAPDH	Hs99999905_m1
TFAP4	Hs00231478_m1
NF-ATc1	Hs00542678_m1
CEBP/beta	Hs00270923_s1
ELK-1	Hs00901847_m1

B. RQ-PCR ChIP/Capture/MeDIP primers.

All Taqman chemistry with 5'FAM/3'TAMRA

Specificity/position	Sequences
Alpha globin promoter chr16:162867-162950	F: GGGCCGGCACTCTTCTG R: GGCCTTGACGTTGGTCTTGT P: CCCACAGACTCAGAGAGAACCCACCATG
Alpha globin exon 1 chr16:163095-163165	F: CCCACCCCTCACTCTGCTT R: TGCGGGAAGTAGGTCTTGGT P: TCCCCGCAGGATGTTCCTGTCC
"359"	F: AAATTCCTAGGCAGGCAGAGATG R: CGACAAGTACAAGGTGAGGGTTT P: CGGGCCACGAGTCCCCAGGT
"218" (LUC7L promoter) chr16:218811-218897	F: GTTCCCCTTGCAATGGTACAG R: GCTCCTTGGGAAGCTACAC P: TTTTCGAGGCCCATCCGGCTC
103 (HS-40) chr16:103394-103470	F: CAGGCTCCAGGCCCATATC R: CCTCCTGCACTGTCCTTTGAC P: TGCCCAAGAGCTCCTTCTGCAACC
"53" Chr16:53195-53285	F: GAGATGCTGGAGTCAGGACCAT R: AGGAGTCAGGAGCAGCAGTCA P: TCCTAGCTCACCCAGAGGCACAGAGG
"82" chr16:82216-82298	F: TCTGCCTGCAAGCATTCG R: TTTCCCAAGTCAGCCGTATGA P: ACCTCCTGGGTGTGACACGCCC
"150" 5' Alpha globin chr16:149736-149822	F: GCACGGCTTGCAGCTACA R: AGATCCTGAGGGCAGCAGATAAG P: CAGGCCCACAGCACCAAGTCTTTTCT
9066: unmethylated (chr 6, 13436197-13436521)	F: AAAGCATCCAAACAGCTGCAT R: TGTCATTGTAGTGCAATCTAAGAATGG P: GACAACAGGGACCGACCAAAT
4994: methylated (chr 22, 48435751- 48436242)	F: CCTTCTTCCTGACCTTGGATGT R: GGGATGTCGAAATGAAAAATGG P: CTGACCGTTTTAACCGAGCTCTCTTTTCGA
6165: methylated (chr 22, 49054726-49055151)	F: ATTGCCGAGTCCACGTAACTG R: TGCTGCAATCCTAGTACATCTCTCA P: ACCAGCCTGGACCGTTCCCACA
11851: methylated (chr 20, 35510497-35510986)	F: AAGACCTCAGTGGCCAACAAA R: ACCTTGCTTCACCCTCACTCAT P: AGACCCACGGAGCTGAAACTCGCC
6583: methylated (chr 22, 30949610- 30950048)	F: TCCCTCTACCCAGATGCTATGATT R: TGGAGAAGTACAGACGTAGAAGTGAAC P: TTTCACCAAAGCCAAAGACAAGGGAACTC
11821: methylated (chr 20, 20574038- 20574521)	F: TGGGGATATTAACAAAGGAAGCA R: CTGTAATACTGTCTCATGTGTCCATCA P: TGTTCCTGGGCTTCTCCCGCTTG
8804: unmethylated (chr 6, 99886966-99887351)	F: GCAGTGCATTATTTTCTTGATCTCA R: AGAAAGAATTTAAGGCAGCTGTATTACTT P: TAATAGGTTAATCCAACAGACTGCATTT
5132: unmethylated (chr 22, 30209493-30209791)	F: TTAACACAAGAGAAACCAAGATTAGTAGCT R: TTGGCAAGTGACAGGCACTTT P: CAGACTTCGCGGGTTCCCAGCCTA

C. Bisulfite sequencing primers.

These are all on Chr 16, positions indicated

Name and position	Type	Sequence
Me1c 132,792-133,261	External Internal	F: TTATGTTTATGTTTGTATTTTATTGG R: CCTAATAAAACAACAACACTCAAC F: TGTTTGTATTTTATTGGTTGTGA R: TCCAAACAAATCCCAAATAA
Me1 133,001-133,700	External Internal	F: TTGGGTTTGTTTTTGGTAGG R: TCCCTTCCTATCCCTAAACCTT F: TGGTTTGGGTTTTAAAGGGTTA R: TTAAACTAACCCCCACCTCAA
Me2 78,001-78,750	External Internal	F: GTTAGGGTGGGGTTGATAGAGATTATTGAG R: CACCAAATCCACTAAAAACAAATAAAACCC F: GTGGGGTTGATAGAGATTATTGAGTTATATT R: AATCCACTAAAAACAAATAAAACCCTATTC
Me3 104001-104600	External Internal	F: AGTGTTTTTGGGATTAGAGGGTTGAT R: CAATCCAAAAATCTCCCCTCCC F: GAGGGTTGATATATGGGGTGTATATAATTG R: CCCCTCCCTTTATTCCATCAAATAT
Me4 158,701-159,100	External Internal	F: GGTTTGGGAGAAAGTTGGTAATTATATTGT R: CATACAACTTCCTCAACTCAAACACATCA F: AAAGTTGGTAATTATATTGTTGGTTATGTTA R: CTCAAACACATCATTAAACATATCATCTAA
Me5 168,001-168,800	External Internal	F: GTATTTATTTAGTTTTTTTGAGGTTGGGG R: CTTCAAATAATCCACCTACCTAAACCTAAC F: TTAGTTTTTTTGAGGTTGGGGAGTT R: TCCACCTACCTAAACCTAACAAAATACTAC
Me6 180,951-181,400	External Internal	F: GTTGTTTGTATGTTTTTAGTTTGAATGG R: CTAATAACTCTCACCTAACAAAAAACCTAT F: TTTAGTTTGAATGGGAAGTGTTAATGTTTT R: CTCACCTAACAAAAAACCTATACCTAAAACT
Me7 103,301-104,000	External Internal	F: GAAGGATGAGAAGTTTTTTTTAGTAGAAGT R: AACATAAAAACAACAATCCTTTCTCTAAC F: GAGAAGTTTTTTTTAGTAGAAGTTATATTGGTT R: CAAAAAATACTACCTTATTATAAATCCATCC
Me8 125,501-126,000	External Internal	F: GATATAAATAGTTTGGAGGAAAATAAGGGA R: ATAAATCTTAAAACAAAACTATAATCACCCTATT F: AATAGTTTGGAGGAAAATAAGGGAGTTT R: TAAAACAAAACTATAATCACCCTATTTACTTCTAT
Me10 94,501-95,500	External Internal	F: GAGTATTTTGTATTGAGGTTGGTGTTTTTTTATG R: CACCTATCTCCTATCACAAACTAAACTAAACCC F: GTATTGAGGTTGGTGTTTTTTTATGAATTAATT R: CACAAACTAAACTAAACCCTCTTAACTAACCC
Me11 150,917-151,399	External Internal	F: TTGGTGTTGTGGGTTTGTT R: AACTCAATATAAATACCCTTCCTAC F: GTTTTGGGGTTGGTGGT R: ATATAAATACCCTTCCTACCCTC

Me12 163,463-163,910	External Internal	F: TTTTTTTTTTAGGGTAGAGGATT R: CACTACCTACAAACCTAACACCT F: TTGTATAGTTTTTAAGTTATTGTTTGT R: CTACAAACCTAACACCTCTCAA
Me13 164,180-164,600	External Internal	F: AAATTAGTAAAATAGGTTGTTTTTAGT R: ACAAAATCCTCTCCACCCTC F: GGTGATTTTTTGAGAGTAGTTTAGTT R: CCCTCCCCAAAACTCAATC
Me14 53,459-53,948	External Internal	F: TAGGTTGTTTGATTGTTTTTTTAGAT R: CCACTCATCCTCTTATTAAACCTTA F: GTTTGATTGTTTTTTTAGATTGTGA R: ATCCTCTTATTAAACCTTACCCAC
Me15 118,507-118,809	External Internal	F: GGTTGTGTTTTTATTTTTTTAAGG R: TAAATCTCACTAATACCACCCATA F: TTTGTAATTTTAGTATTTTGGGAG R: ATACATCAAATCCACTCTCTATATCT
Me16 120,314-120,634	External Internal	F: AGTTTGAAGTTGGAATAGGTGTT R: CAAAATTCACCCATATTCATTC F: GTTGGAATAGGTGTTTAAATTATTT R: AACCCTATTTTCTTAACCCTTC
Me18 170,141-170,542	External Internal	F: TGGTATGGTGGTATGTATTTGTG R: CACCTTTCCAAAACCTCTATC F: GGGAGGTTGAGGTAGGAGAA R: ACCCAACTTCTTCCACAAAAC
Me21 105,973-106,310	External Internal	F: GGTTTGATTTTTATTTTTTATTTTTG R: CATTTTCATAACATCTTTATATTTCTTAC F: GTTGAGAATAAGTTTGGGGGA R: CAAAAAAATAAACAAAATCCATAAC
Me22 107,794-108,293	External Internal	F: GGGTTTAGGTGAGGTAGGTTAGTA R: ATATCATAACACTAAAAAACTCATCTTA F: AGGTGAGGTAGGTTAGTAAGTTTT R: CCTACCAATCAACTCTCTACATACA
Me23 158,395-158,780	External Internal	F: AGGTAGAGTTTGGTGTGAAAGTAGT R: AAAACTAAAAACTAAATCCAAAATAA F: GGATATTTTATGGGTATAAGTTGTT R: ATTCTTACCTCTCCAAAACCTC
Me24 159,009-159,554	External Internal	F: AGATTAAGAAATAATGTAAGTAGGTGG R: TACAATTCTCCCTCCCCAAC F: TGATATGTTTAATGATGTGTTTGAG R: CCCCAACCCCATACAACTAC
Me28 164,964-165,243	External Internal	F: TAAGGAGGTAGAGGTTGTAGTGAG R: TCTCCCCAAACCCTTACTCA F: GGTAGAGGTTGTAGTGAGTTAAGATT R: CAAACCCTTACTCAAACCCTAC
Me29 166,986-167,541	External Internal	F: TGTTTTTTTTTATTATTAAGATTTATTTT R: CTTTTCCTACCCTATCCAACTAC F: GGTTTTGTTTAGGTTAAGGGTTA R: AAACTCAAACACACACAAACTACC
Me30	External	F: GTGGTTTTTGAATAAAGTTTGAGT

167,518-168,154	Internal	R: CAATAAAACAAACTCCCCAAC F: GGTAGTTTGTGTGTGTTTGAGTT R: AACTAAACCCTAAACTAAACCCA
Alpha 1 exon 1-2 162,899-163,204	External Internal	F: TTTGGTTTTTATAGATTTAGAGAGAAT R: AAATCACCAACAAACAATAACTTA F: GAGAGAATTTATTATGGTGTTGTTT R: TAACCCTTAACCTAAACAAAACC
HBAIs 166,150-166,719	External Internal	F: TGGAGAGTAGGAGGGGTTTTG R: AAAAAACAACACCATAATAAATTCTCT F: TTTTTGAGTTTTTATGGGTTAG R: CCATAATAAATTCTCTCTAAATCTATA
HBA1a 165,755-166,167	External Internal	F: GTGTGTATTTGTGTGATGGTTAGA R: CCCTAACCCATAAAAACTCAA F: TAGAGGAGGTTGAATGTTTTAG R: CCCATAAAAACTCAAAAAAAT

D. EMSA oligos.

CGs are highlighted in yellow and mutated bases in fuchsia.

Name	Sequence 5' → 3'
Sp1	F: CAGGCCCCGCCCGGACT R: AGTCCCGGGCGGGGCCTG
Me13.2	F: GGAGCAGAGCACCTTG***CG***GCCTTACTCCTGCA R: GGTGCAGGAGTAAGGCCGCAAGGTGCTCTGCT
Me15.5	F: GGGGCGTGGTGGCAGT***CG***CCTGTAGTCCCAGC R: GGGCTGGGACTACAGGCGACTGCCACCACGCC
Me16.2	F: GGGAGCACCCACTAAA***CG***GAAAAAAACAAACA R: GGTGTTTGTTTTTTTCCGTTTAGTGGGTGCTC
Me18.5	F: GGGTGAGCCGGAATGG***CG***CCACTGCACTCACC R: GGGGTGAGTGCAGTGGCGCCATTCCGGCTCAC
Me21.5	F: GGTGGGGCTGTGGCTG***CG***TTTCATCAACTGTC R: GGGACAGTTGATGAAACGCAGCCACAGCCCCA
Me30.2	F: GGGTGCCAGGCATGGG***CG***TGGACAGCAGCTGG R: GGCCAGCTGCTGTCCACGCCCATGCCTGGCAC
16.2_mut1	F: GGGAGAACCCACTAAA***CG***GAAAAAAACAAACA R: TGTTTGTTTTTTTCCGTTTAGTGGGTTCTCCC
16.2_mut2	F: GGGAGCACACTAAA***CG***GAAAAAAACAAACA R: TGTTTGTTTTTTTCCGTTTAGTGTGTGCTCCC
16.2_mut3	F: GGGAGCACCCACTGAA***CG***GAAAAAAACAAACA R: TGTTTGTTTTTTTCCGTTCAGTGGGTGCTCCC
16.2_mut4	F: GGGAGCACCCACTAAA***TG***GAAAAAAACAAACA R: TGTTTGTTTTTTTCCATTTAGTGGGTGCTCCC
16.2_mut5	F: GGGAGCACCCACTAAA***CG***GAGAAAAAACAAACA R: TGTTTGTTTTTCTCCGTTTAGTGGGTGCTCCC
16.2_mut6	F: GGGAGCACCCACTAAA***CG***GAAATAAACAAACA R: TGTTTGTTTATTTCCGTTTAGTGGGTGCTCCC
16.2_mut7	F: GGGAGCACCCACTAAA***CG***GAAAAAAGCAAACA R: TGTTTGCTTTTTTCCGTTTAGTGGGTGCTCCC

16.2_mut8	F: GGGAGCACCCACTAAA***CG***GAAAAAAAACAGACA R: TGTCTGTTTTTTTCCGTTTAGTGGGTGCTCCC

E. RPA (Random primer amplification) primers for capture protocol.
Acu-1 recognition site in red. N= random A, T, C, or G

RG3A: GTT TCC CAG CTG AAG TCN NNN NN
RG4: GTT TCC CAG CTG AAG TC

F. Antibodies used for chromatin immunoprecipitation
H3K4me3 (Millipore 05-745R), 5µl
H3K27me3 (Millipore 07-449), 10µl
H3Ac (Millipore 06-599), 10µl
H3K9me3 (Abcam Ab8898), 10µl
H4Ac (Millipore 06-866) 10µl

G. Antibody used for MeDIP immunoprecipitation
Mouse monoclonal against 5-methylcytosine (Diagenode)

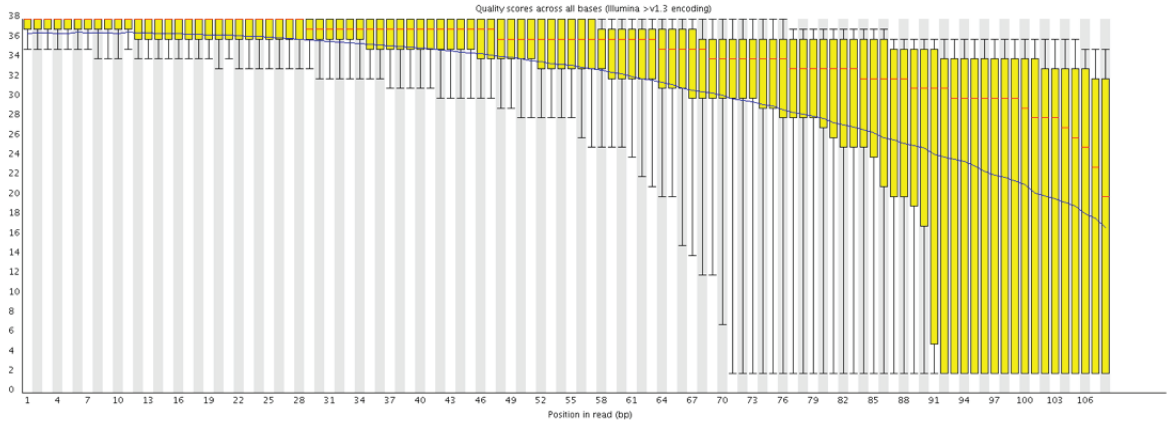
H. DNA affinity chromatography oligos.
The spacer is coloured red and the Pst1 recognition site is green

MS16.2G_F:
biotin-5' CAAAACTGCAGGAGCACCCACTAAACGGAAAAAAACAAACA3'

MS16.2G_R
5' TGTTTGTTTTTTTCCGTTTAGTGGGTGCTCCTGCAGTTTTG3'

Appendix 2: Supplementary Figures

S4.1A. N1 per base sequence quality



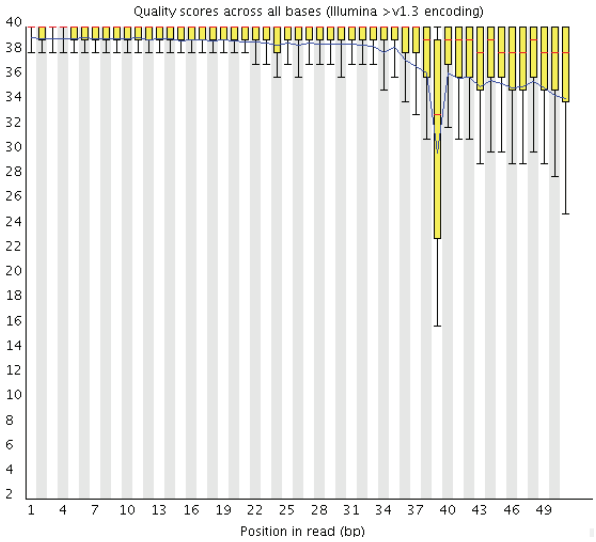
S4.1B. N1 over-represented sequence

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CCTGCTTTTCGGGAAAACGGGATCACCACGATGGAACAGGTTAACGCAGGAATGCGCGTAGCCCGTCGGCAGAATCGACCATTTCTGCCATCACCCGGGCAGTTTGTT	46653	0.12928918384695803	No Hit

S4.1C. CLL1 per base sequence quality

Forward



Reverse

Per base sequence quality

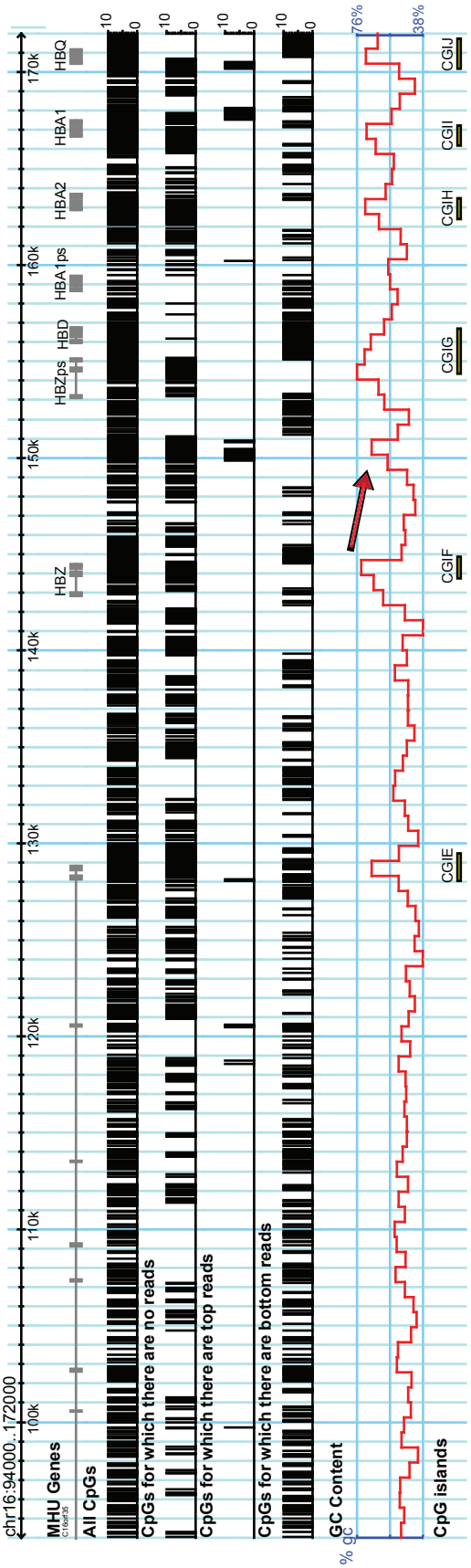


S4.1D CLL1 over-represented sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT	674654	2.479437220885804	Illumina Single End PCR Primer 1 (100% over 51bp)
GATCGGAAGAGCGTCGTGTAGGGAAAGAGGGTAGATCTCGGTGGTCGCCGT	324594	1.1929232544033042	Illumina Single End PCR Primer 1 (98% over 51bp)
GATCGGAAGAGCGTCGTGTAGGGGAAGAGGGTAGATCTCGGTGGTCGCCGT	243916	0.8964215867238344	Illumina Single End PCR Primer 1 (96% over 29bp)
CGGCATTCCTGCTGAACCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG	243143	0.8935807157414571	Illumina Single End PCR Primer 1 (100% over 32bp)
GATCGGAAGAGCGGCGGGTAGGGGAAGAGGGGAGATCTCGGGGGGCGCCGG	227574	0.8363627075595282	No Hit
GATCGGAAGAGCGGCGGGGAGGGGAAGAGGGGAGATCTCGGGGGGCGCCGG	225872	0.8301076462244622	No Hit
GATCGGAAGAGGGGCGGGGGGGGGAAGAGGGGAGATCTCGGGGGGGGCCGG	194488	0.714767549315113	No Hit

Figure S4.1 Quality control from N1 and CLL1. **S4.1**A shows the quality of the reads for sample N1, showing poor quality beyond 50nt. **S4.1B** shows an over-represented sequence which did not match any known genome. **S4.1C** shows quality scores for sample CLL1, showing that the forward reads have good quality but the reverse reads are unsuable. This is due to dye quenching from repeated sequencing of adapters, present in the overrepresented sequences. **S4.1D** shows that the over-represented sequences in the reverse reads are also all versions of the adapters.



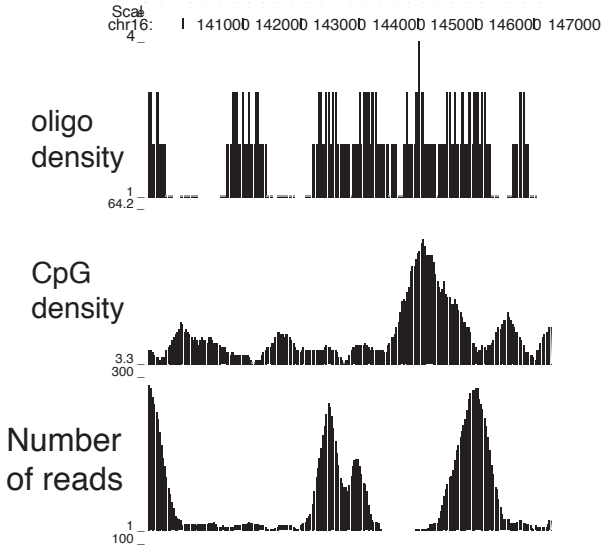
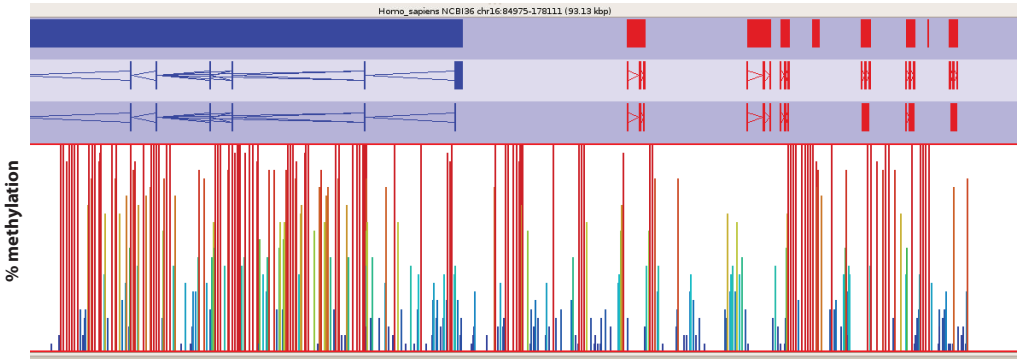


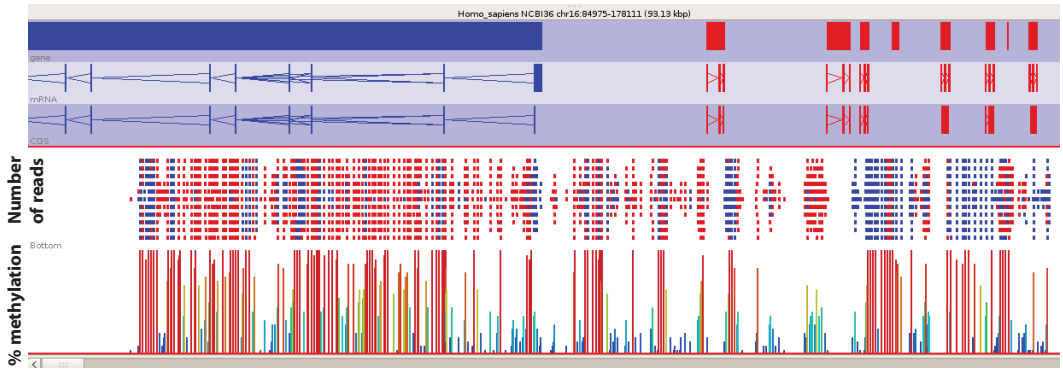
Figure S4.2 B. Comparison between density of capture oligos in the top track, CpG density in the middle track and number of reads in the bottom track. The CpG dense region pictured here, although covered by oligos, is associated with a gap in reads coming from that region.

Figure S4.2 A. GBrowse tracks showing first a spike for every CpG in the capture area, then a spike for every CpG which did not yield results by TBS, then a spike for every CpG which had a top read and finally a spike for every CpG which had a bottom read. It can be seen that CpGs for which there are top reads and no reads do tend to be concentrated in regions of high GC density. However this is not the case for every CGI, as CGI-G is well represented by bottom reads. One cluster of top reads and no reads (150K region) falls in a region that, although not technically a CGI, does have a high GC content, red arrow. These observations do suggest that poor capture of CpGs in GC rich regions could result from lack of complete denaturation.

S4.3A. seqmonk snapshot of target region showing percentage methylation



S4.3B. seqmonk snapshot of target region showing percentage methylation and nuber of reads



S4.3C. Histogram of frequency of number of reads

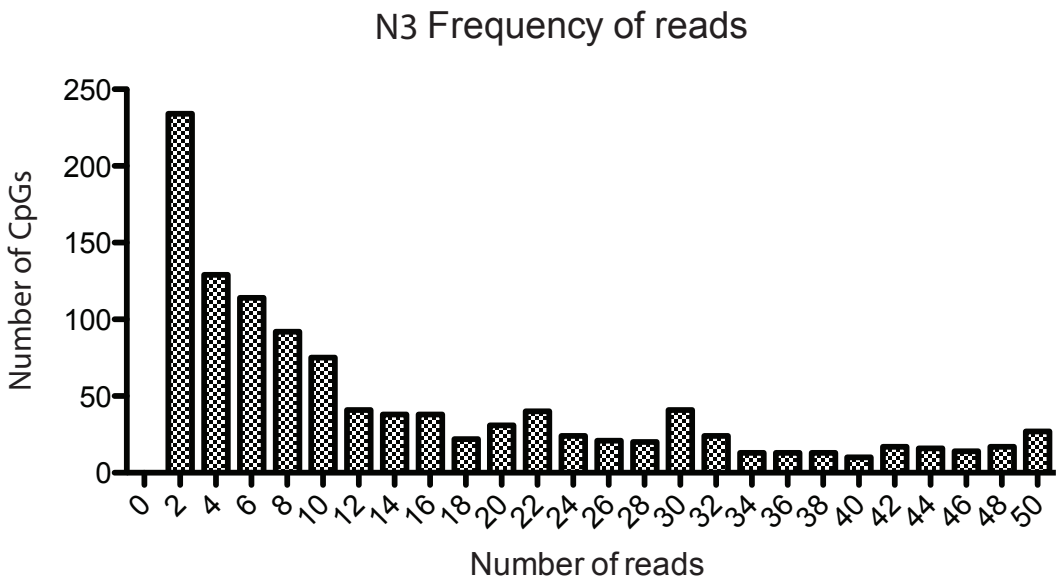
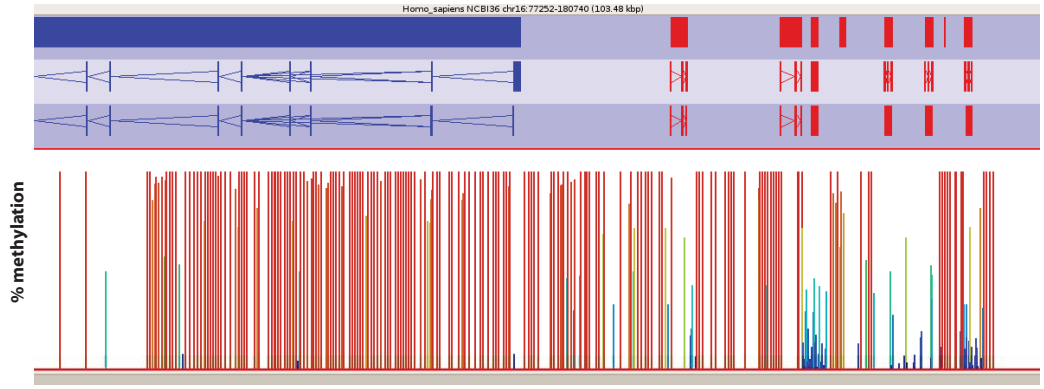
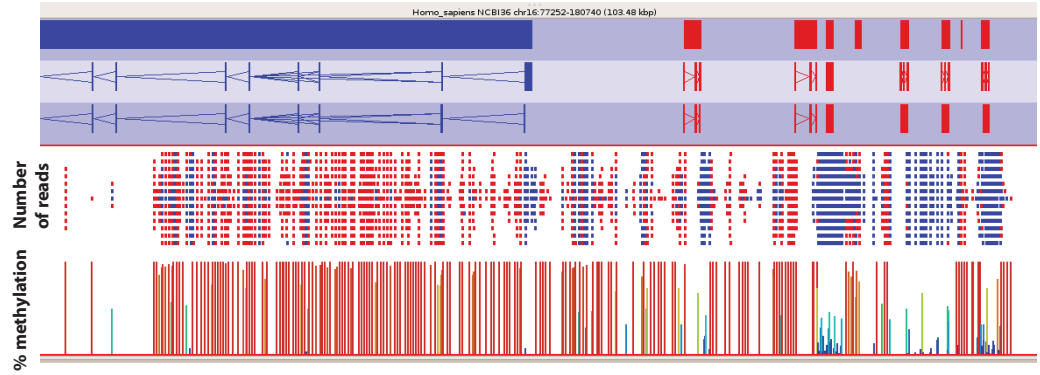


Figure S4.3 Seqmonk snapshots and histogram of frequency of reads for N3. **A.** Sequmonk track showing methylation % of individual CpGs. This is represented graphically both by the height of the peak and its colour (red= fully methylated, blue= unmethylated). **B**. Seqmonk track including the number of reads in the top of the panel. **C.** Histogram showing by how many reads are CpGs covered. In this example, many CpGs are covered by relatively few reads. CpGs covered by fewer than 10 reads were not used for analysis.

S4.4A. seqmonk snapshot of target region showing percentage methylation



S4.4B. seqmonk snapshot of target region showing percentage methylation and nuber of reads



S4.4C. Histogram of frequency of number of reads

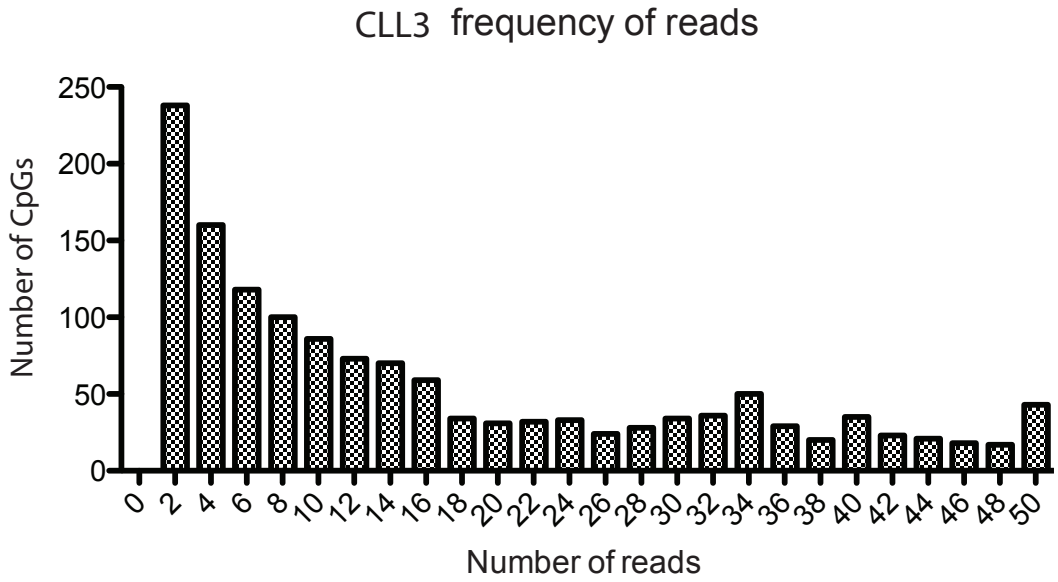
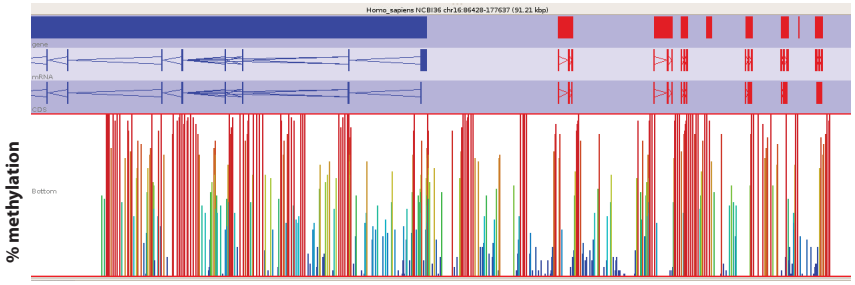
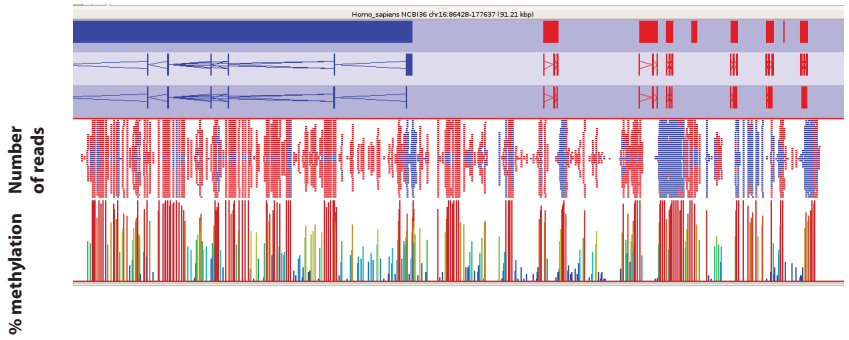


Figure S4.4 Seqmonk snapshots and histogram of frequency of reads for CLL3. A. Sequmonk track showing methylation % of individual CpGs. This is represented graphically both by the height of the peak and its colour (red= fully methylated, blue= unmethylated). **B.** Seqmonk track including the number of reads in the top of the panel. **C.** Histogram showing by how many reads are CpGs covered. In this example, many CpGs are covered by relatively few reads. CpGs covered by fewer than 10 reads were not used for analysis.

S4.5A. seqmonk snapshot of target region showing percentage methylation



S4.5B. seqmonk snapshot of target region showing percentage methylation and nuber of reads



S4.5C. Histogram of frequency of number of reads

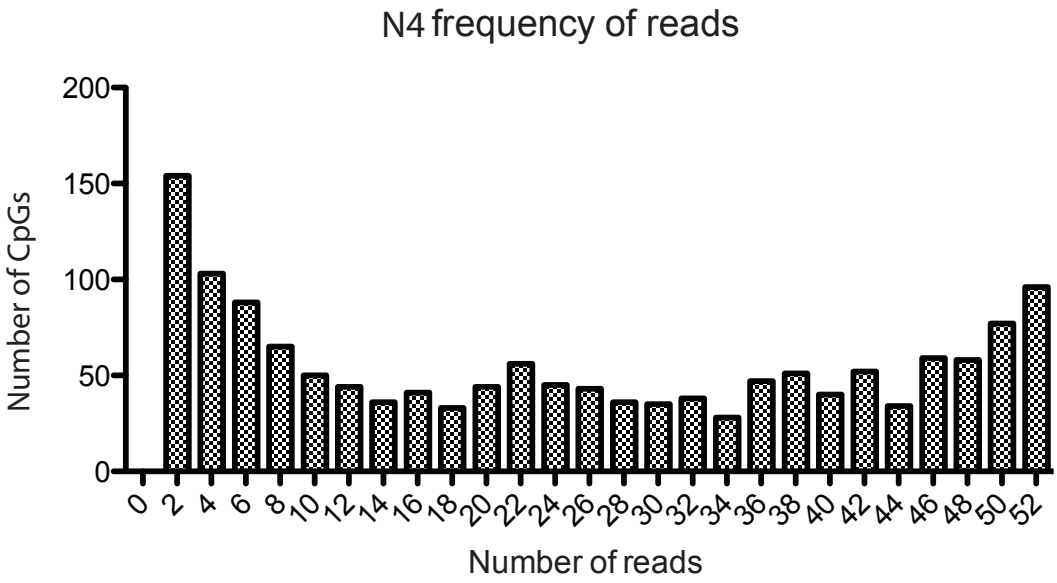
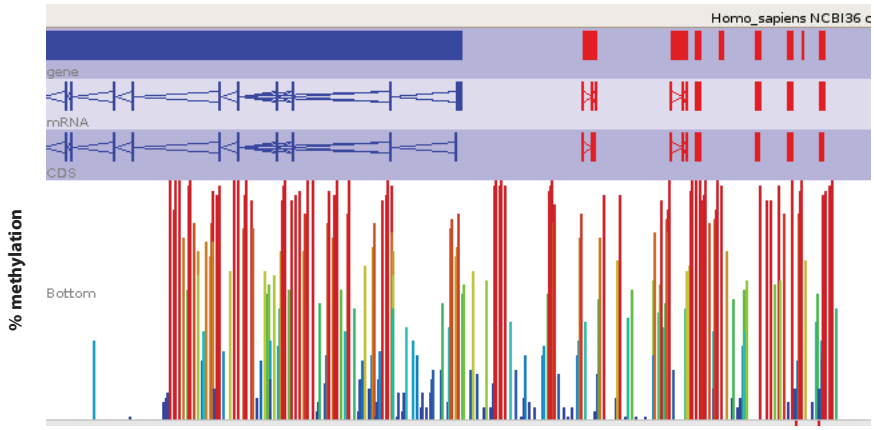
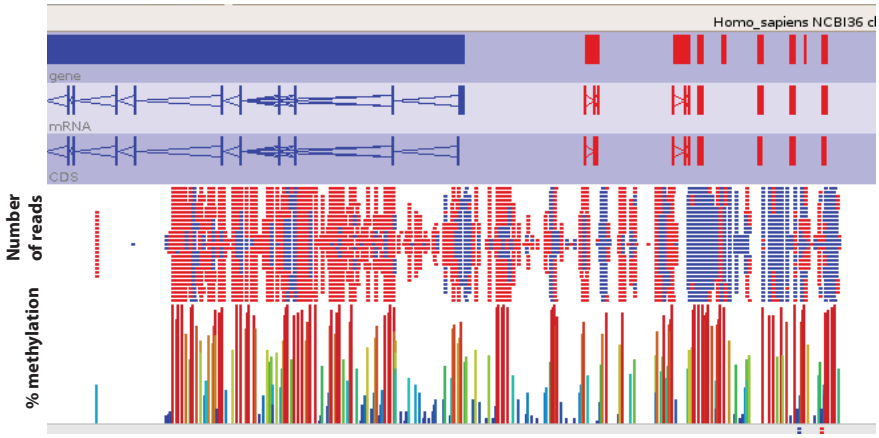


Figure S4.5 Seqmonk snapshots and histogram of frequency of reads for N4. A. Sequmonk track showing methylation % of individual CpGs. This is represented graphically both by the height of the peak and its colour (red= fully methylated, blue= unmethylated). **B.** Seqmonk track including the number of reads in the top of the panel. **C.** Histogram showing by how many reads are CpGs covered. Note this sample was prepared with the optimised protocol and it is clear that fewer of the CpGs are covered by low numbers of reads while the number of CpGs covered by numerous reads has increased.

S4.6A. seqmonk snapshot of target region showing percentage methylation



S4.6B. seqmonk snapshot of target region showing percentage methylation and nuber of reads



S4.6C. Histogram of frequency of number of reads

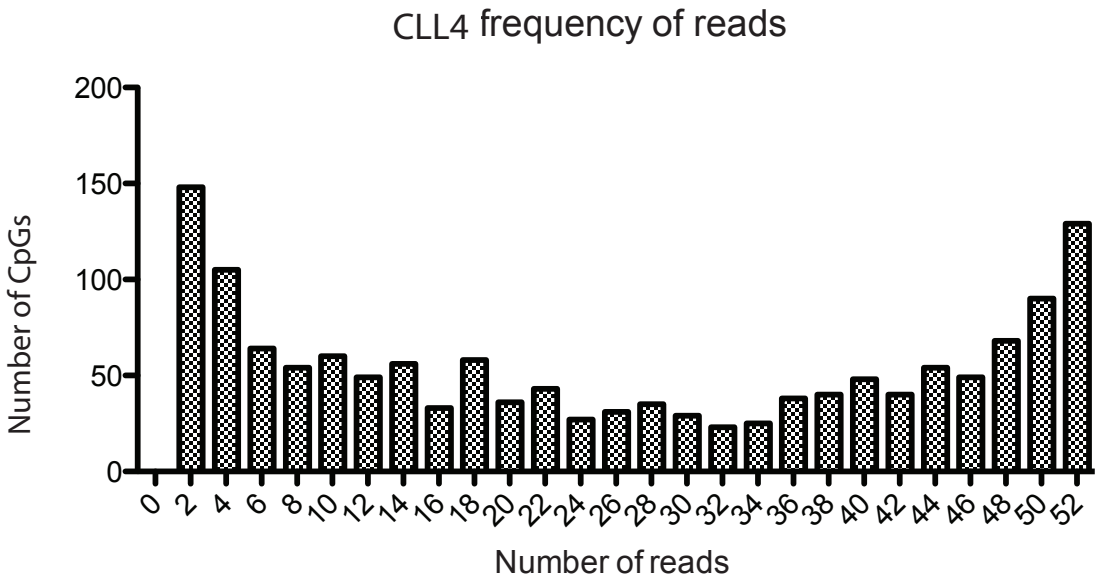


Figure S4.6 Seqmonk snapshots and histogram of frequency of reads for CLL4. A. Sequmonk track showing methylation % of individual CpGs. This is represented graphically both by the height of the peak and its colour (red= fully methylated, blue= unmethylated). **B**. Seqmonk track including the number of reads in the top of the panel. **C.** Histogram showing by how many reads are CpGs covered. Note this sample was prepared with the optimised protocol and it is clear that fewer of the CpGs are covered by low numbers of reads while the number of CpGs covered by numerous reads has increased.

References

ANDO, Y., and Y. HAYASHIZAKI, 2006 Restriction landmark genomic scanning. Nat Protoc **1:** 2774-2783.

ARANBURU, A., R. CARLSSON, C. PERSSON and T. LEANDERSON, 2001 Transcription factor AP-4 is a ligand for immunoglobulin-kappa promoter E-box elements. The Biochemical journal **354:** 431-438.

ARITA, K., M. ARIYOSHI, H. TOCHIO, Y. NAKAMURA and M. SHIRAKAWA, 2008 Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature **455:** 818-821.

BACHMAN, K. E., M. R. ROUNTREE and S. B. BAYLIN, 2001 Dnmt3a and Dnmt3b are transcriptional repressors that exhibit unique localization properties to heterochromatin. The Journal of biological chemistry **276:** 32282-32287.

BALL, M. P., J. B. LI, Y. GAO, J. H. LEE, E. M. LEPROUST *et al.*, 2009 Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nature biotechnology **27:** 361-368.

BARRETO, G., A. SCHAFER, J. MARHOLD, D. STACH, S. K. SWAMINATHAN *et al.*, 2007 Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. Nature **445:** 671-675.

BASU-ROY, U., E. SEO, L. RAMANATHAPURAM, T. B. RAPP, J. A. PERRY *et al.*, 2011 Sox2 maintains self renewal of tumor-initiating cells in osteosarcomas. Oncogene.

BAYLIN, S. B., and P. A. JONES, 2011 A decade of exploring the cancer epigenome - biological and translational implications. Nature reviews. Cancer **11:** 726-734.

BECK, S., and V. K. RAKYAN, 2008 The methylome: approaches for global DNA methylation profiling. Trends Genet **24:** 231-237.

BEDFORD, M. T., and P. D. VAN HELDEN, 1987 Hypomethylation of DNA in pathological conditions of the human prostate. Cancer research **47:** 5274-5276.

BENNETT, L. B., K. H. TAYLOR, G. L. ARTHUR, F. B. RAHMATPANAH, S. I. HOOSHMAND *et al.*, 2010 Epigenetic regulation of WNT signaling in chronic lymphocytic leukemia. Epigenomics **2:** 53-70.

BERTILACCIO, M. T., C. SCIELZO, M. MUZIO and F. CALIGARIS-CAPPIO, 2010 An overview of chronic lymphocytic leukaemia biology. Best practice & research. Clinical haematology **23:** 21-32.

BESTOR, T., A. LAUDANO, R. MATTALIANO and V. INGRAM, 1988 Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of molecular biology **203:** 971-983.

BESTOR, T. H., 2000 The DNA methyltransferases of mammals. Human molecular genetics **9:** 2395-2402.

BILLOT, K., J. SOEUR, F. CHEREAU, I. ARROUSS, H. MERLE-BERAL *et al.*, 2011 Deregulation of Aiolos expression in chronic lymphocytic leukemia is associated with epigenetic modifications. Blood **117:** 1917-1927.

BIRD, A., 2002 DNA methylation patterns and epigenetic memory. Genes Dev **16:** 6-21.

BIRD, A., M. TAGGART, M. FROMMER, O. J. MILLER and D. MACLEOD, 1985 A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell **40:** 91-99.

BLUM, K. A., Z. LIU, D. M. LUCAS, P. CHEN, Z. XIE *et al.*, 2010 Phase I trial of low dose decitabine targeting DNA hypermethylation in patients with chronic lymphocytic leukaemia and non-Hodgkin lymphoma: dose-limiting myelosuppression without evidence of DNA hypomethylation. British journal of haematology **150:** 189-195.

BOCK, C., E. M. TOMAZOU, A. B. BRINKMAN, F. MULLER, F. SIMMER *et al.*, 2010 Quantitative comparison of genome-wide DNA methylation mapping technologies. Nature biotechnology **28:** 1106-1114.

BOISE, L. H., B. PETRYNIAK, X. MAO, C. H. JUNE, C. Y. WANG *et al.*, 1993 The NFAT-1 DNA binding complex in activated T cells contains Fra-1 and JunB. Molecular and cellular biology **13:** 1911-1919.

BOROWCZYK, E., K. N. MOHAN, L. D'AIUTO, M. C. CIRIO and J. R. CHAILLET, 2009 Identification of a region of the DNMT1 methyltransferase that regulates the maintenance of genomic imprints. Proceedings of the National Academy of Sciences of the United States of America **106:** 20806-20811.

BOSTICK, M., J. K. KIM, P. O. ESTEVE, A. CLARK, S. PRADHAN *et al.*, 2007 UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science **317:** 1760-1764.

BOULTWOOD, J., and J. S. WAINSCOAT, 2007 Gene silencing by DNA methylation in haematological malignancies. Br J Haematol **138:** 3-11.

BOUMBER, Y. A., Y. KONDO, X. CHEN, L. SHEN, Y. GUO *et al.*, 2008 An Sp1/Sp3 binding polymorphism confers methylation protection. PLoS Genet **4:** e1000162.

BOURC'HIS, D., G. L. XU, C. S. LIN, B. BOLLMAN and T. H. BESTOR, 2001 Dnmt3L and the establishment of maternal genomic imprints. Science **294:** 2536-2539.

BRANDEIS, M., D. FRANK, I. KESHET, Z. SIEGFRIED, M. MENDELSOHN *et al.*, 1994 Sp1 elements protect a CpG island from de novo methylation. Nature **371:** 435-438.

BRINKMAN, A. B., F. SIMMER, K. MA, A. KAAN, J. ZHU *et al.*, 2010 Whole-genome DNA methylation profiling using MethylCap-seq. Methods **52:** 232-236.

BRUNIQUEL, D., and R. H. SCHWARTZ, 2003 Selective, stable demethylation of the interleukin-2 gene enhances transcription by an active process. Nature immunology **4:** 235-240.

BYRD, J. C., G. MARCUCCI, M. R. PARTHUN, J. J. XIAO, R. B. KLISOVIC *et al.*, 2005 A phase 1 and pharmacodynamic study of depsipeptide (FK228) in chronic lymphocytic leukemia and acute myeloid leukemia. Blood **105:** 959-967.

CALIN, G. A., C. D. DUMITRU, M. SHIMIZU, R. BICHI, S. ZUPO *et al.*, 2002 Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proceedings of the National Academy of Sciences of the United States of America **99:** 15524-15529.

CAMERON, E. E., K. E. BACHMAN, S. MYOHANEN, J. G. HERMAN and S. B. BAYLIN, 1999 Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer. Nature genetics **21:** 103-107.

CEDAR, H., and Y. BERGMAN, 2009a Linking DNA methylation and histone modification: patterns and paradigms. Nature reviews. Genetics **10:** 295-304.

CEDAR, H., and Y. BERGMAN, 2009b Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet **10:** 295-304.

CHALITCHAGORN, K., S. SHUANGSHOTI, N. HOURPAI, N. KONGRUTTANACHOK, P. TANGKIJVANICH *et al.*, 2004 Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. Oncogene **23:** 8841-8846.

CHAO, W., and P. A. D'AMORE, 2008a IGF2: epigenetic regulation and role in development and disease. Cytokine & growth factor reviews **19:** 111-120.

CHAO, W., and P. A. D'AMORE, 2008b IGF2: epigenetic regulation and role in development and disease. Cytokine Growth Factor Rev **19:** 111-120.

CHEN, P. Y., S. J. COKUS and M. PELLEGRINI, 2010 BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics **11:** 203.

CHEN, S. S., A. RAVAL, A. J. JOHNSON, E. HERTLEIN, T. H. LIU *et al.*, 2009 Epigenetic changes during disease progression in a murine model of human chronic lymphocytic leukemia. Proc Natl Acad Sci U S A **106:** 13433-13438.

CHEN, T., Y. UEDA, J. E. DODGE, Z. WANG and E. LI, 2003 Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. Molecular and cellular biology **23:** 5594-5605.

CHEN, Z. X., and A. D. RIGGS, 2011 DNA methylation and demethylation in mammals. The Journal of biological chemistry **286:** 18347-18353.

CHINNAPPAN, D., D. XIAO, A. RATNASARI, C. ANDRY, T. C. KING *et al.*, 2009a Transcription factor YY1 expression in human gastrointestinal cancer cells. International journal of oncology **34:** 1417-1423.

CHINNAPPAN, D., D. XIAO, A. RATNASARI, C. ANDRY, T. C. KING *et al.*, 2009b Transcription factor YY1 expression in human gastrointestinal cancer cells. Int J Oncol **34:** 1417-1423.

CHIORAZZI, N., and M. FERRARINI, 2011 Cellular origin(s) of chronic lymphocytic leukemia: cautionary notes and additional considerations and possibilities. Blood **117:** 1781-1791.

CHODAVARAPU, R. K., S. FENG, Y. V. BERNATAVICHUTE, P. Y. CHEN, H. STROUD *et al.*, 2010 Relationship between nucleosome positioning and DNA methylation. Nature **466:** 388-392.

CHOI, S. H., S. WORSWICK, H. M. BYUN, T. SHEAR, J. C. SOUSSA *et al.*, 2009 Changes in DNA methylation of tandem DNA repeats are different from interspersed repeats in cancer. Int J Cancer **125:** 723-729.

CHUANG, L. S., H. I. IAN, T. W. KOH, H. H. NG, G. XU *et al.*, 1997 Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science **277:** 1996-2000.

CLARK, S. J., J. HARRISON and M. FROMMER, 1995 CpNpG methylation in mammalian cells. Nat Genet **10:** 20-27.

COKUS, S. J., S. FENG, X. ZHANG, Z. CHEN, B. MERRIMAN *et al.*, 2008 Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature **452:** 215-219.

CONERLY, M. L., S. S. TEVES, D. DIOLAITI, M. ULRICH, R. N. EISENMAN *et al.*, 2010 Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis. Genome research **20:** 1383-1390.

COOLEN, M. W., C. STIRZAKER, J. Z. SONG, A. L. STATHAM, Z. KASSIR *et al.*, 2010 Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. Nature cell biology **12:** 235-246.

COOPER, D. N., and M. KRAWCZAK, 1989 Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Human genetics **83:** 181-188.

CORTELLINO, S., J. XU, M. SANNAI, R. MOORE, E. CARETTI *et al.*, 2011 Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. Cell **146:** 67-79.

COTNOIR-WHITE, D., D. LAPERRIERE and S. MADER, 2011 Evolution of the repertoire of nuclear receptor binding sites in genomes. Molecular and cellular endocrinology **334:** 76-82.

CROSSLEY, M., E. WHITELAW, A. PERKINS, G. WILLIAMS, Y. FUJIWARA *et al.*, 1996 Isolation and characterization of the cDNA encoding BKLF/TEF-2, a major CACCC-box-binding protein in erythroid cells and selected other cells. Molecular and cellular biology **16:** 1695-1705.

CRUICKSHANK, M. N., E. FENWICK, M. KARIMI, L. J. ABRAHAM and D. ULGIATI, 2009 Cell- and stage-specific chromatin structure across the Complement receptor 2 (CR2/CD21) promoter coincide with CBF1 and C/EBP-beta binding in B cells. Molecular immunology **46:** 2613-2622.

DAMESHEK, W., 1967 Chronic lymphocytic leukemia--an accumulative disease of immunolgically incompetent lymphocytes. Blood **29:** Suppl:566-584.

DAMLE, R. N., F. GHIOTTO, A. VALETTO, E. ALBESIANO, F. FAIS *et al.*, 2002 B-cell chronic lymphocytic leukemia cells express a surface membrane phenotype of activated, antigen-experienced B lymphocytes. Blood **99:** 4087-4093.

DAMMANN, R. H., S. KIRSCH, U. SCHAGDARSURENGIN, T. DANSRANJAVIN, E. GRADHAND *et al.*, 2010 Frequent aberrant methylation of the imprinted IGF2/H19 locus and LINE1 hypomethylation in ovarian carcinoma. International journal of oncology **36:** 171-179.

DASKALOS, A., G. NIKOLAIDIS, G. XINARIANOS, P. SAVVARI, A. CASSIDY *et al.*, 2009 Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. International journal of cancer. Journal international du cancer **124:** 81-87.

DE CARVALHO, D. D., J. S. YOU and P. A. JONES, 2010 DNA methylation and cellular reprogramming. Trends in cell biology **20:** 609-617.

DE MARZO, A. M., V. L. MARCHI, E. S. YANG, R. VEERASWAMY, X. LIN *et al.*, 1999 Abnormal regulation of DNA methyltransferase expression during colorectal carcinogenesis. Cancer research **59:** 3855-3860.

DE SMET, C., and A. LORIOT, 2010 DNA hypomethylation in cancer: Epigenetic scars of a neoplastic journey. Epigenetics **5**.

DEBATIN, K. M., 2007 Chronic lymphocytic leukemia: keeping cell death at bay. Cell **129:** 853-855.

DELHOMMEAU, F., S. DUPONT, V. DELLA VALLE, C. JAMES, S. TRANNOY *et al.*, 2009 Mutation in TET2 in myeloid cancers. The New England journal of medicine **360:** 2289-2301.

DENG, J., R. SHOEMAKER, B. XIE, A. GORE, E. M. LEPROUST *et al.*, 2009 Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol **27:** 353-360.

DESAI, S., and J. PINILLA-IBARZ, 2012 Front-line therapy for chronic lymphocytic leukemia. Cancer control : journal of the Moffitt Cancer Center **19:** 26-36.

DI CROCE, L., V. A. RAKER, M. CORSARO, F. FAZI, M. FANELLI *et al.*, 2002 Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor. Science **295:** 1079-1082.

DIGNAM, J. D., R. M. LEBOVITZ and R. G. ROEDER, 1983 Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res **11:** 1475-1489.

DING, F., and J. R. CHAILLET, 2002 In vivo stabilization of the Dnmt1 (cytosine-5)-methyltransferase protein. Proceedings of the National Academy of Sciences of the United States of America **99:** 14861-14866.

DODGE, J. E., M. OKANO, F. DICK, N. TSUJIMOTO, T. CHEN *et al.*, 2005 Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. J Biol Chem **280:** 17986-17991.

DOHNER, H., S. STILGENBAUER, A. BENNER, E. LEUPOLT, A. KROBER *et al.*, 2000 Genomic aberrations and survival in chronic lymphocytic leukemia. The New England journal of medicine **343:** 1910-1916.

DOI, A., I. H. PARK, B. WEN, P. MURAKAMI, M. J. ARYEE *et al.*, 2009 Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet **41:** 1350-1353.

DOWN, T. A., V. K. RAKYAN, D. J. TURNER, P. FLICEK, H. LI *et al.*, 2008 A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol **26:** 779-785.

DUBOVSKY, J. A., D. WANG, J. J. POWERS, E. BERCHMANS, M. A. SMITH *et al.*, 2011 Restoring the functional immunogenicity of chronic lymphocytic leukemia using epigenetic modifiers. Leukemia research **35:** 394-404.

DUPONT, J. M., J. TOST, H. JAMMES and I. G. GUT, 2004 De novo quantitative bisulfite sequencing using the pyrosequencing technology. Analytical biochemistry **333:** 119-127.

EADS, C. A., K. D. DANENBERG, K. KAWAKAMI, L. B. SALTZ, C. BLAKE *et al.*, 2000 MethyLight: a high-throughput assay to measure DNA methylation. Nucleic acids research **28:** E32.

EADS, C. A., K. D. DANENBERG, K. KAWAKAMI, L. B. SALTZ, P. V. DANENBERG *et al.*, 1999 CpG island hypermethylation in human colorectal tumors is not associated with DNA methyltransferase overexpression. Cancer research **59:** 2302-2306.

ECKHARDT, F., J. LEWIN, R. CORTESE, V. K. RAKYAN, J. ATTWOOD *et al.*, 2006a DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet **38:** 1378-1385.

ECKHARDT, F., J. LEWIN, R. CORTESE, V. K. RAKYAN, J. ATTWOOD *et al.*, 2006b DNA methylation profiling of human chromosomes 6, 20 and 22. Nature genetics **38:** 1378-1385.

EDEN, A., F. GAUDET, A. WAGHMARE and R. JAENISCH, 2003 Chromosomal instability and tumors promoted by DNA hypomethylation. Science **300:** 455.

EDWARDS, J. R., A. H. O'DONNELL, R. A. ROLLINS, H. E. PECKHAM, C. LEE *et al.*, 2010 Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome research **20:** 972-980.

EGGER, G., G. LIANG, A. APARICIO and P. A. JONES, 2004 Epigenetics in human disease and prospects for epigenetic therapy. Nature **429:** 457-463.

EHRLICH, M., 2002 DNA methylation in cancer: too much, but also too little. Oncogene **21:** 5400-5413.

EHRLICH, M., 2009 DNA hypomethylation in cancer cells. Epigenomics **1:** 239-259.

EHRLICH, M., M. A. GAMA-SOSA, L. H. HUANG, R. M. MIDGETT, K. C. KUO *et al.*, 1982 Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic acids research **10:** 2709-2721.

EL-DEIRY, W. S., B. D. NELKIN, P. CELANO, R. W. YEN, J. P. FALCO *et al.*, 1991 High expression of the DNA methyltransferase gene characterizes human neoplastic cells and progression stages of colon cancer. Proceedings of the National Academy of Sciences of the United States of America **88:** 3470-3474.

EL-MAARRI, O., M. S. KARETA, T. MIKESKA, T. BECKER, A. DIAZ-LACAVA *et al.*, 2009 A systematic search for DNA methyltransferase polymorphisms reveals a rare DNMT3L variant associated with subtelomeric hypomethylation. Human molecular genetics **18:** 1755-1768.

ELGIN, S. C., 1988 The formation and function of DNase I hypersensitive sites in the process of gene activation. The Journal of biological chemistry **263:** 19259-19262.

ESPADA, J., E. BALLESTAR, M. F. FRAGA, A. VILLAR-GAREA, A. JUARRANZ *et al.*, 2004 Human DNA methyltransferase 1 is required for maintenance of the histone H3 modification pattern. The Journal of biological chemistry **279:** 37175-37184.

ESTELLER, M., G. GAIDANO, S. N. GOODMAN, V. ZAGONEL, D. CAPELLO *et al.*, 2002 Hypermethylation of the DNA repair gene O(6)-methylguanine DNA methyltransferase and survival of patients with diffuse large B-cell lymphoma. Journal of the National Cancer Institute **94:** 26-32.

FABIANI, E., G. LEONE, M. GIACHELIA, F. D'ALO, M. GRECO *et al.*, 2010 Analysis of genome-wide methylation and gene expression induced by 5-aza-2'-deoxycytidine identifies BCL2L10 as a frequent methylation target in acute myeloid leukemia. Leukemia & lymphoma **51:** 2275-2284.

FABRIS, S., V. BOLLATI, L. AGNELLI, F. MORABITO, V. MOTTA *et al.*, 2011 Biological and clinical relevance of quantitative global methylation of repetitive DNA sequences in chronic lymphocytic leukemia. Epigenetics : official journal of the DNA Methylation Society **6:** 188-194.

FACKELMAYER, F. O., K. DAHM, A. RENZ, U. RAMSPERGER and A. RICHTER, 1994 Nucleic-acid-binding properties of hnRNP-U/SAF-A, a nuclear-matrix protein which binds DNA and RNA in vivo and in vitro. European journal of biochemistry / FEBS **221:** 749-757.

FATEMI, M., A. HERMANN, H. GOWHER and A. JELTSCH, 2002 Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. European journal of biochemistry / FEBS **269:** 4981-4984.

FATEMI, M., and P. A. WADE, 2006 MBD family proteins: reading the epigenetic code. J Cell Sci **119:** 3033-3037.

FEINBERG, A. P., C. W. GEHRKE, K. C. KUO and M. EHRLICH, 1988 Reduced genomic 5-methylcytosine content in human colonic neoplasia. Cancer research **48:** 1159-1161.

FELDMAN, N., A. GERSON, J. FANG, E. LI, Y. ZHANG *et al.*, 2006 G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. Nat Cell Biol **8:** 188-194.

FELTUS, F. A., E. K. LEE, J. F. COSTELLO, C. PLASS and P. M. VERTINO, 2006 DNA motifs associated with aberrant CpG island methylation. Genomics **87:** 572-579.

FENAUX, P., G. J. MUFTI, E. HELLSTROM-LINDBERG, V. SANTINI, C. FINELLI *et al.*, 2009 Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study. The lancet oncology **10:** 223-232.

FENG, S., S. J. COKUS, X. ZHANG, P. Y. CHEN, M. BOSTICK *et al.*, Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A **107:** 8689-8694.

FILION, G. J., S. ZHENILO, S. SALOZHIN, D. YAMADA, E. PROKHORTCHOUK *et al.*, 2006 A family of human zinc finger proteins that bind methylated DNA and repress transcription. Molecular and cellular biology **26:** 169-181.

FLATAU, E., E. BOGENMANN and P. A. JONES, 1983 Variable 5-methylcytosine levels in human tumor cell lines and fresh pediatric tumor explants. Cancer research **43:** 4901-4905.

FLUSBERG, B. A., D. R. WEBSTER, J. H. LEE, K. J. TRAVERS, E. C. OLIVARES *et al.*, 2010 Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature methods **7:** 461-465.

FRAGA, M. F., E. BALLESTAR, G. MONTOYA, P. TAYSAVANG, P. A. WADE *et al.*, 2003 The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. Nucleic acids research **31:** 1765-1774.

FRAGA, M. F., E. BALLESTAR, M. F. PAZ, S. ROPERO, F. SETIEN *et al.*, 2005a Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences of the United States of America **102:** 10604-10609.

FRAGA, M. F., E. BALLESTAR, A. VILLAR-GAREA, M. BOIX-CHORNET, J. ESPADA *et al.*, 2005b Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nature genetics **37:** 391-400.

FRANCIS, N. J., R. E. KINGSTON and C. L. WOODCOCK, 2004 Chromatin compaction by a polycomb group protein complex. Science **306:** 1574-1577.

FROMMER, M., L. E. MCDONALD, D. S. MILLAR, C. M. COLLIS, F. WATT *et al.*, 1992 A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A **89:** 1827-1831.

FUJII, S., R. Z. LUO, J. YUAN, M. KADOTA, M. OSHIMURA *et al.*, 2003 Reactivation of the silenced and imprinted alleles of ARHI is associated with increased

histone H3 acetylation and decreased histone H3 lysine 9 methylation. Human molecular genetics **12:** 1791-1800.

FUJITA, N., N. SHIMOTAKE, I. OHKI, T. CHIBA, H. SAYA *et al.*, 2000 Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. Mol Cell Biol **20:** 5107-5118.

FUJITA, N., S. WATANABE, T. ICHIMURA, S. TSURUZOE, Y. SHINKAI *et al.*, 2003 Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. The Journal of biological chemistry **278:** 24132-24138.

FULOP, Z., B. CSERNUS, B. TIMAR, A. SZEPESI and A. MATOLCSY, 2003 Microsatellite instability and hMLH1 promoter hypermethylation in Richter's transformation of chronic lymphocytic leukemia. Leukemia **17:** 411-415.

GALM, O., J. G. HERMAN and S. B. BAYLIN, 2006 The fundamental role of epigenetics in hematopoietic malignancies. Blood Rev **20:** 1-13.

GAMA-SOSA, M. A., V. A. SLAGEL, R. W. TREWYN, R. OXENHANDLER, K. C. KUO *et al.*, 1983a The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res **11:** 6883-6894.

GAMA-SOSA, M. A., R. Y. WANG, K. C. KUO, C. W. GEHRKE and M. EHRLICH, 1983b The 5-methylcytosine content of highly repeated sequences in human DNA. Nucleic Acids Res **11:** 3087-3095.

GAUDET, F., J. G. HODGSON, A. EDEN, L. JACKSON-GRUSBY, J. DAUSMAN *et al.*, 2003 Induction of tumors in mice by genomic hypomethylation. Science **300:** 489-492.

GEORGEL, P. T., R. A. HOROWITZ-SCHERER, N. ADKINS, C. L. WOODCOCK, P. A. WADE *et al.*, 2003 Chromatin compaction by human MeCP2. Assembly of novel secondary chromatin structures in the absence of DNA methylation. The Journal of biological chemistry **278:** 32181-32188.

GILBERT, N., and B. RAMSAHOYE, 2005 The relationship between chromatin structure and transcriptional activity in mammalian genomes. Briefings in functional genomics & proteomics **4:** 129-142.

GILBERT, N., I. THOMSON, S. BOYLE, J. ALLAN, B. RAMSAHOYE *et al.*, 2007 DNA methylation affects nuclear organization, histone modifications, and linker histone binding but not chromatin compaction. The Journal of cell biology **177:** 401-411.

GIMENEZ, J., C. MONTGIRAUD, J. P. PICHON, B. BONNAUD, M. ARSAC *et al.*, 2010 Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic acids research **38:** 2229-2246.

GONG, Z., and J. K. ZHU, 2011 Active DNA demethylation by oxidation and repair. Cell research.

GORE, S. D., S. BAYLIN, E. SUGAR, H. CARRAWAY, C. B. MILLER *et al.*, 2006 Combined DNA methyltransferase and histone deacetylase inhibition in the treatment of myeloid neoplasms. Cancer research **66:** 6361-6369.

GOWHER, H., O. LEISMANN and A. JELTSCH, 2000 DNA of Drosophila melanogaster contains 5-methylcytosine. The EMBO journal **19:** 6918-6923.

GRATIO, V., F. WALKER, T. LEHY, M. LABURTHE and D. DARMOUL, 2009 Aberrant expression of proteinase-activated receptor 4 promotes colon cancer cell proliferation through a persistent signaling that involves Src and ErbB-2

kinase. International journal of cancer. Journal international du cancer **124:** 1517-1525.

GU, H., C. BOCK, T. S. MIKKELSEN, N. JAGER, Z. D. SMITH *et al.*, 2010 Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. Nature methods **7:** 133-136.

GUO, J. U., D. K. MA, H. MO, M. P. BALL, M. H. JANG *et al.*, 2011 Neuronal activity modifies the DNA methylation landscape in the adult brain. Nature neuroscience **14:** 1345-1351.

GUPTA, R., A. NAGARAJAN and N. WAJAPEYEE, 2010 Advances in genome-wide DNA methylation analysis. BioTechniques **49:** iii-xi.

HAGEMANN, S., O. HEIL, F. LYKO and B. BRUECKNER, 2011 Azacytidine and decitabine induce gene-specific and non-random DNA demethylation in human cancer cell lines. PLoS ONE **6:** e17388.

HAHN, M. A., X. WU, A. X. LI, T. HAHN and G. P. PFEIFER, 2011 Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. PLoS ONE **6:** e18844.

HAJKOVA, P., S. J. JEFFRIES, C. LEE, N. MILLER, S. P. JACKSON *et al.*, 2010 Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science **329:** 78-82.

HALLEK, M., K. FISCHER, G. FINGERLE-ROWSON, A. M. FINK, R. BUSCH *et al.*, 2010 Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. Lancet **376:** 1164-1174.

HANADA, M., D. DELIA, A. AIELLO, E. STADTMAUER and J. C. REED, 1993 bcl-2 gene hypomethylation and high-level expression in B-cell chronic lymphocytic leukemia. Blood **82:** 1820-1828.

HANDA, V., and A. JELTSCH, 2005 Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. Journal of molecular biology **348:** 1103-1112.

HANSEN, K. D., W. TIMP, H. C. BRAVO, S. SABUNCIYAN, B. LANGMEAD *et al.*, 2011 Increased methylation variation in epigenetic domains across cancer types. Nature genetics **43:** 768-775.

HARRIS, E. Y., N. PONTS, A. LEVCHUK, K. L. ROCH and S. LONARDI, 2010 BRAT: bisulfite-treated reads analysis tool. Bioinformatics **26:** 572-573.

HATTORI, N., T. ABE, M. SUZUKI, T. MATSUYAMA, S. YOSHIDA *et al.*, 2004 Preference of DNA methyltransferases for CpG islands in mouse embryonic stem cells. Genome research **14:** 1733-1740.

HAYASHIZAKI, Y., S. HIROTSUNE, Y. OKAZAKI, I. HATADA, H. SHIBATA *et al.*, 1993 Restriction landmark genomic scanning method and its various applications. Electrophoresis **14:** 251-258.

HE, Y. F., B. Z. LI, Z. LI, P. LIU, Y. WANG *et al.*, 2011 Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science **333:** 1303-1307.

HELLMAN, A., and A. CHESS, 2007 Gene body-specific methylation on the active X chromosome. Science **315:** 1141-1143.

HENDRICH, B., C. ABBOTT, H. MCQUEEN, D. CHAMBERS, S. CROSS *et al.*, 1999a Genomic structure and chromosomal mapping of the murine and human Mbd1, Mbd2, Mbd3, and Mbd4 genes. Mammalian genome : official journal of the International Mammalian Genome Society **10:** 906-912.

HENDRICH, B., U. HARDELAND, H. H. NG, J. JIRICNY and A. BIRD, 1999b The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature **401:** 301-304.

HENIKOFF, S., and K. AHMAD, 2005 Assembly of variant histones into chromatin. Annual review of cell and developmental biology **21:** 133-153.

HERMAN, J. G., J. R. GRAFF, S. MYOHANEN, B. D. NELKIN and S. B. BAYLIN, 1996 Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proceedings of the National Academy of Sciences of the United States of America **93:** 9821-9826.

HERMAN, J. G., A. UMAR, K. POLYAK, J. R. GRAFF, N. AHUJA *et al.*, 1998 Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proceedings of the National Academy of Sciences of the United States of America **95:** 6870-6875.

HERNANDEZ-BLAZQUEZ, F. J., M. HABIB, J. M. DUMOLLARD, C. BARTHELEMY, M. BENCHAIB *et al.*, 2000 Evaluation of global DNA hypomethylation in human colon cancer tissues by immunohistochemistry and image analysis. Gut **47:** 689-693.

HERVOUET, E., L. LALIER, E. DEBIEN, M. CHERAY, A. GEAIRON *et al.*, 2010a Disruption of Dnmt1/PCNA/UHRF1 interactions promotes tumorigenesis from human and mice glial cells. PLoS ONE **5:** e11333.

HERVOUET, E., F. M. VALLETTE and P. F. CARTRON, 2009 Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. Epigenetics : official journal of the DNA Methylation Society **4:** 487-499.

HERVOUET, E., F. M. VALLETTE and P. F. CARTRON, 2010b Dnmt1/Transcription factor interactions: an alternative mechanism of DNA methylation inheritance. Genes & cancer **1:** 434-443.

HITE, K. C., V. H. ADAMS and J. C. HANSEN, 2009 Recent advances in MeCP2 structure and function. Biochemistry and cell biology = Biochimie et biologie cellulaire **87:** 219-227.

HODGES, E., A. D. SMITH, J. KENDALL, Z. XUAN, K. RAVI *et al.*, 2009 High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res **19:** 1593-1605.

HOGLUND, A., L. M. NILSSON, S. V. MURALIDHARAN, L. A. HASVOLD, P. MERTA *et al.*, 2011 Therapeutic implications for the induced levels of chk1 in myc-expressing cancer cells. Clinical cancer research : an official journal of the American Association for Cancer Research **17:** 7067-7079.

HUANG, Y., W. A. PASTOR, Y. SHEN, M. TAHILIANI, D. R. LIU *et al.*, 2010 The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS One **5:** e8888.

IKEGAMI, K., M. IWATANI, M. SUZUKI, M. TACHIBANA, Y. SHINKAI *et al.*, 2007 Genome-wide and locus-specific DNA hypomethylation in G9a deficient mouse embryonic stem cells. Genes to cells : devoted to molecular & cellular mechanisms **12:** 1-11.

ILLINGWORTH, R., A. KERR, D. DESOUSA, H. JORGENSEN, P. ELLIS *et al.*, 2008 A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol **6:** e22.

ILLINGWORTH, R. S., and A. P. BIRD, 2009 CpG islands--'a rough guide'. FEBS Lett **583:** 1713-1720.

IMAMURA, T., J. OHGANE, S. ITO, T. OGAWA, N. HATTORI *et al.*, 2001 CpG island of rat sphingosine kinase-1 gene: tissue-dependent DNA methylation status and multiple alternative first exons. Genomics **76:** 117-125.

INOUE, A., and Y. ZHANG, 2011 Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. Science **334:** 194.

INOUE, S., R. T. SNOWDEN, M. J. DYER and G. M. COHEN, 2004 CDDO induces apoptosis via the intrinsic pathway in lymphoid cells. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K **18:** 948-952.

IQBAL, K., S. G. JIN, G. P. PFEIFER and P. E. SZABO, 2011 Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proceedings of the National Academy of Sciences of the United States of America **108:** 3642-3647.

IRIZARRY, R. A., C. LADD-ACOSTA, B. CARVALHO, H. WU, S. A. BRANDENBURG *et al.*, 2008a Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res **18:** 780-790.

IRIZARRY, R. A., C. LADD-ACOSTA, B. CARVALHO, H. WU, S. A. BRANDENBURG *et al.*, 2008b Comprehensive high-throughput arrays for relative methylation (CHARM). Genome research **18:** 780-790.

IRIZARRY, R. A., C. LADD-ACOSTA, B. WEN, Z. WU, C. MONTANO *et al.*, 2009 The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet.

ITO, S., A. C. D'ALESSIO, O. V. TARANOVA, K. HONG, L. C. SOWERS *et al.*, 2010 Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature **466:** 1129-1133.

ITO, S., L. SHEN, Q. DAI, S. C. WU, L. B. COLLINS *et al.*, 2011 Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science **333:** 1300-1303.

JACKSON-GRUSBY, L., C. BEARD, R. POSSEMATO, M. TUDOR, D. FAMBROUGH *et al.*, 2001 Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. Nature genetics **27:** 31-39.

JAHNER, D., H. STUHLMANN, C. L. STEWART, K. HARBERS, J. LOHLER *et al.*, 1982 De novo methylation and expression of retroviral genomes during mouse embryogenesis. Nature **298:** 623-628.

JENUWEIN, T., and C. D. ALLIS, 2001 Translating the histone code. Science **293:** 1074-1080.

JEONG, S., G. LIANG, S. SHARMA, J. C. LIN, S. H. CHOI *et al.*, 2009a Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. Mol Cell Biol **29:** 5366-5376.

JEONG, S., G. LIANG, S. SHARMA, J. C. LIN, S. H. CHOI *et al.*, 2009b Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. Molecular and cellular biology **29:** 5366-5376.

JIA, D., R. Z. JURKOWSKA, X. ZHANG, A. JELTSCH and X. CHENG, 2007 Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature **449:** 248-251.

JIN, F., S. C. DOWDY, Y. XIONG, N. L. EBERHARDT, K. C. PODRATZ *et al.*, 2005 Up-regulation of DNA methyltransferase 3B expression in endometrial cancers. Gynecologic oncology **96:** 531-538.

JIN, S. G., C. GUO and G. P. PFEIFER, 2008 GADD45A does not promote DNA demethylation. PLoS genetics **4:** e1000013.

JONES, P. A., and S. B. BAYLIN, 2007 The epigenomics of cancer. Cell **128:** 683-692.

JONES, P. A., and G. LIANG, 2009 Rethinking how DNA methylation patterns are maintained. Nature reviews. Genetics **10:** 805-811.

JONES, P. L., G. J. VEENSTRA, P. A. WADE, D. VERMAAK, S. U. KASS *et al.*, 1998 Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nature genetics **19:** 187-191.

JUNG, P., A. MENSSEN, D. MAYR and H. HERMEKING, 2008 AP4 encodes a c-MYC-inducible repressor of p21. Proceedings of the National Academy of Sciences of the United States of America **105:** 15046-15051.

KANAI, Y., and S. HIROHASHI, 2007 Alterations of DNA methylation associated with abnormalities of DNA methyltransferases in human cancers during transition from a precancerous to a malignant state. Carcinogenesis **28:** 2434-2442.

KANDURI, M., N. CAHILL, H. GORANSSON, C. ENSTROM, F. RYAN *et al.*, 2010 Differential genome-wide array-based methylation profiles in prognostic subsets of chronic lymphocytic leukemia. Blood **115:** 296-305.

KANEDA, M., M. OKANO, K. HATA, T. SADO, N. TSUJIMOTO *et al.*, 2004 Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature **429:** 900-903.

KAPLAN, D., D. SMITH, H. MEYERSON, N. PECORA and K. LEWANDOWSKA, 2001 CD5 expression by B lymphocytes and its regulation upon Epstein-Barr virus transformation. Proceedings of the National Academy of Sciences of the United States of America **98:** 13850-13853.

KATO, Y., M. KANEDA, K. HATA, K. KUMAKI, M. HISANO *et al.*, 2007 Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Human molecular genetics **16:** 2272-2280.

KESHET, I., Y. SCHLESINGER, S. FARKASH, E. RAND, M. HECHT *et al.*, 2006 Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat Genet **38:** 149-153.

KHULAN, B., R. F. THOMPSON, K. YE, M. J. FAZZARI, M. SUZUKI *et al.*, 2006 Comparative isoschizomer profiling of cytosine methylation: the HELP assay. Genome Res **16:** 1046-1055.

KIM, M. S., T. KONDO, I. TAKADA, M. Y. YOUN, Y. YAMAMOTO *et al.*, 2009 DNA demethylation in hormone-induced transcriptional derepression. Nature **461:** 1007-1012.

KIM, Y. I., A. GIULIANO, K. D. HATCH, A. SCHNEIDER, M. A. NOUR *et al.*, 1994 Global DNA hypomethylation increases progressively in cervical dysplasia and carcinoma. Cancer **74:** 893-899.

KIRILLOV, A., B. KISTLER, R. MOSTOSLAVSKY, H. CEDAR, T. WIRTH *et al.*, 1996 A role for nuclear NF-kappaB in B-cell-specific demethylation of the Igkappa locus. Nature genetics **13:** 435-441.

KLEIN, U., Y. TU, G. A. STOLOVITZKY, M. MATTIOLI, G. CATTORETTI *et al.*, 2001 Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. The Journal of experimental medicine **194:** 1625-1638.

KLOSE, R. J., S. A. SARRAF, L. SCHMIEDEBERG, S. M. MCDERMOTT, I. STANCHEVA *et al.*, 2005 DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. Molecular cell **19:** 667-678.

KLUG, M., S. HEINZ, C. GEBHARD, L. SCHWARZFISCHER, S. W. KRAUSE *et al.*, 2010 Active DNA demethylation in human postmitotic cells correlates with activating histone modifications, but not transcription levels. Genome biology **11:** R63.

KN, H., S. BASSAL, C. TIKELLIS and A. EL-OSTA, 2004 Expression analysis of the epigenetic methyltransferases and methyl-CpG binding protein families in the normal B-cell and B-cell chronic lymphocytic leukemia (CLL). Cancer biology & therapy **3:** 989-994.

KNAPEN, D., L. VERGAUWEN, K. LAUKENS and R. BLUST, 2009 Best practices for hybridization design in two-colour microarray analysis. Trends in biotechnology **27:** 406-414.

KOH, K. P., A. YABUUCHI, S. RAO, Y. HUANG, K. CUNNIFF *et al.*, 2011 Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell stem cell **8:** 200-213.

KOK, R. M., D. E. SMITH, R. BARTO, A. M. SPIJKERMAN, T. TEERLINK *et al.*, 2007 Global DNA methylation measured by liquid chromatography-tandem mass spectrometry: analytical technique, reference values and determinants in healthy subjects. Clinical chemistry and laboratory medicine : CCLM / FESCC **45:** 903-911.

KOUZARIDES, T., 2007 Chromatin modifications and their function. Cell **128:** 693-705.

KRIAUCIONIS, S., and N. HEINTZ, 2009 The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science **324:** 929-930.

KRUEGER, F., and S. R. ANDREWS, 2011 Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics **27:** 1571-1572.

LAIRD, C. D., N. D. PLEASANT, A. D. CLARK, J. L. SNEEDEN, K. M. HASSAN *et al.*, 2004a Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. Proc Natl Acad Sci U S A **101:** 204-209.

LAIRD, C. D., N. D. PLEASANT, A. D. CLARK, J. L. SNEEDEN, K. M. HASSAN *et al.*, 2004b Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. Proceedings of the National Academy of Sciences of the United States of America **101:** 204-209.

LANASA, M. C., 2010 Novel insights into the biology of CLL. Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program **2010:** 70-76.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LANGMEAD, B., C. TRAPNELL, M. POP and S. L. SALZBERG, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology **10:** R25.

LAURENT, L., E. WONG, G. LI, T. HUYNH, A. TSIRIGOS *et al.*, 2010 Dynamic changes in the human methylome during differentiation. Genome Res **20:** 320-331.

LEE, H. S., B. H. KIM, N. Y. CHO, E. J. YOO, M. CHOI *et al.*, 2009 Prognostic implications of and relationship between CpG island hypermethylation and repetitive DNA hypomethylation in hepatocellular carcinoma. Clin Cancer Res **15:** 812-820.

LEE, P. J., L. L. WASHER, D. J. LAW, C. R. BOLAND, I. L. HORON *et al.*, 1996 Limited up-regulation of DNA methyltransferase in human colon cancer reflecting increased cell proliferation. Proceedings of the National Academy of Sciences of the United States of America **93:** 10366-10370.

LEE, T. F., J. ZHAI and B. C. MEYERS, Conservation and divergence in eukaryotic DNA methylation. Proc Natl Acad Sci U S A **107:** 9027-9028.

LEHMANN, U., K. BRAKENSIEK and H. KREIPE, 2004 Role of epigenetic changes in hematological malignancies. Ann Hematol **83:** 137-152.

LEHNERTZ, B., Y. UEDA, A. A. DERIJCK, U. BRAUNSCHWEIG, L. PEREZ-BURGOS *et al.*, 2003a Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Current biology : CB **13:** 1192-1200.

LEHNERTZ, B., Y. UEDA, A. A. DERIJCK, U. BRAUNSCHWEIG, L. PEREZ-BURGOS *et al.*, 2003b Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol **13:** 1192-1200.

LEONHARDT, H., A. W. PAGE, H. U. WEIER and T. H. BESTOR, 1992 A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell **71:** 865-873.

LEY, T. J., L. DING, M. J. WALTER, M. D. MCLELLAN, T. LAMPRECHT *et al.*, 2010 DNMT3A mutations in acute myeloid leukemia. The New England journal of medicine **363:** 2424-2433.

LI, E., C. BEARD, A. C. FORSTER, T. H. BESTOR and R. JAENISCH, 1993 DNA methylation, genomic imprinting, and mammalian development. Cold Spring Harbor symposia on quantitative biology **58:** 297-305.

LI, R., C. YU, Y. LI, T. W. LAM, S. M. YIU *et al.*, 2009 SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics **25:** 1966-1967.

LIANG, G., M. F. CHAN, Y. TOMIGAHARA, Y. C. TSAI, F. A. GONZALES *et al.*, 2002 Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Molecular and cellular biology **22:** 480-491.

LIEB, J. D., 2003 Genome-wide mapping of protein-DNA interactions by chromatin immunoprecipitation and DNA microarray hybridization. Methods in molecular biology **224:** 99-109.

LIENERT, F., C. WIRBELAUER, I. SOM, A. DEAN, F. MOHN *et al.*, 2011 Identification of genetic elements that autonomously determine DNA methylation states. Nature genetics **43:** 1091-1097.

LIN, C. H., B. LI, S. SWANSON, Y. ZHANG, L. FLORENS *et al.*, 2008 Heterochromatin protein 1a stimulates histone H3 lysine 36 demethylation by the Drosophila KDM4A demethylase. Molecular cell **32:** 696-706.

LIN, I. G., and C. L. HSIEH, 2001 Chromosomal DNA demethylation specified by protein binding. EMBO reports **2:** 108-112.

LIN, I. G., T. J. TOMZYNSKI, Q. OU and C. L. HSIEH, 2000 Modulation of DNA binding protein affinity directly affects target site demethylation. Molecular and cellular biology **20:** 2343-2349.

LISTER, R., and J. R. ECKER, 2009 Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res **19:** 959-966.

LISTER, R., M. PELIZZOLA, R. H. DOWEN, R. D. HAWKINS, G. HON *et al.*, 2009 Human DNA methylomes at base resolution show widespread epigenomic differences. Nature **462:** 315-322.

LOPEZ-SERRA, L., E. BALLESTAR, M. F. FRAGA, M. ALAMINOS, F. SETIEN *et al.*, 2006 A profile of methyl-CpG binding domain protein occupancy of hypermethylated promoter CpG islands of tumor suppressor genes in human cancer. Cancer research **66:** 8342-8346.

LORINCZ, M. C., D. R. DICKERSON, M. SCHMITT and M. GROUDINE, 2004 Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat Struct Mol Biol **11:** 1068-1075.

LORIOT, A., E. DE PLAEN, T. BOON and C. DE SMET, 2006 Transient down-regulation of DNMT1 methyltransferase leads to activation and stable hypomethylation of MAGE-A1 in melanoma cells. The Journal of biological chemistry **281:** 10118-10126.

LORIOT, A., C. STERPIN, O. DE BACKER and C. DE SMET, 2008 Mouse embryonic stem cells induce targeted DNA demethylation within human MAGE-A1 transgenes. Epigenetics : official journal of the DNA Methylation Society **3:** 38-42.

LORSBACH, R. B., J. MOORE, S. MATHEW, S. C. RAIMONDI, S. T. MUKATIRA *et al.*, 2003 TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K **17:** 637-641.

LUEDI, P. P., F. S. DIETRICH, J. R. WEIDMAN, J. M. BOSKO, R. L. JIRTLE *et al.*, 2007 Computational and experimental identification of novel human imprinted genes. Genome research **17:** 1723-1730.

LYKO, F., D. STACH, A. BRENNER, S. STILGENBAUER, H. DOHNER *et al.*, 2004 Quantitative analysis of DNA methylation in chronic lymphocytic leukemia patients. Electrophoresis **25:** 1530-1535.

LYNCH, M. D., A. J. SMITH, M. DE GOBBI, M. FLENLEY, J. R. HUGHES *et al.*, 2011 An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. The EMBO journal.

MAITI, A., and A. C. DROHAT, 2011 Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. The Journal of biological chemistry **286:** 35334-35338.

MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER *et al.*, 2010 Target-enrichment strategies for next-generation sequencing. Nature methods **7:** 111-118.

MARQUES, M., L. LAFLAMME, A. L. GERVAIS and L. GAUDREAU, 2010 Reconciling the positive and negative roles of histone H2A.Z in gene transcription. Epigenetics : official journal of the DNA Methylation Society **5:** 267-272.

MATSUO, K., J. SILKE, O. GEORGIEV, P. MARTI, N. GIOVANNINI *et al.*, 1998 An embryonic demethylation mechanism involving binding of transcription factors to replicating DNA. The EMBO journal **17:** 1446-1453.

MAUNAKEA, A. K., R. P. NAGARAJAN, M. BILENKY, T. J. BALLINGER, C. D'SOUZA *et al.*, Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature **466:** 253-257.

MEEHAN, R. R., J. D. LEWIS and A. P. BIRD, 1992 Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. Nucleic acids research **20:** 5085-5092.

MEISSNER, A., A. GNIRKE, G. W. BELL, B. RAMSAHOYE, E. S. LANDER *et al.*, 2005 Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic acids research **33:** 5868-5877.

MEISSNER, A., T. S. MIKKELSEN, H. GU, M. WERNIG, J. HANNA *et al.*, 2008 Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature **454:** 766-770.

MELKI, J. R., and S. J. CLARK, 2002 DNA methylation changes in leukaemia. Seminars in cancer biology **12:** 347-357.

MELKI, J. R., P. C. VINCENT and S. J. CLARK, 1999 Concurrent DNA hypermethylation of multiple genes in acute myeloid leukemia. Cancer Res **59:** 3730-3740.

MELKI, J. R., P. WARNECKE, P. C. VINCENT and S. J. CLARK, 1998 Increased DNA methyltransferase expression in leukaemia. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K **12:** 311-316.

MESSMER, B. T., D. MESSMER, S. L. ALLEN, J. E. KOLITZ, P. KUDALKAR *et al.*, 2005 In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. The Journal of clinical investigation **115:** 755-764.

METIVIER, R., R. GALLAIS, C. TIFFOCHE, C. LE PERON, R. Z. JURKOWSKA *et al.*, 2008 Cyclical DNA methylation of a transcriptionally active promoter. Nature **452:** 45-50.

MILLER, C. A., and J. D. SWEATT, 2007 Covalent modification of DNA regulates memory formation. Neuron **53:** 857-869.

MILNE, T. A., M. E. MARTIN, H. W. BROCK, R. K. SLANY and J. L. HESS, 2005 Leukemogenic MLL fusion proteins bind across a broad region of the Hox a9 locus, promoting transcription and multiple histone modifications. Cancer research **65:** 11367-11374.

MIRSAIDOV, U. M., D. WANG, W. TIMP and G. TIMP, 2010 Molecular diagnostics for personal medicine using a nanopore. Wiley interdisciplinary reviews. Nanomedicine and nanobiotechnology **2:** 367-381.

MITTLER, G., F. BUTTER and M. MANN, 2009a A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome research **19:** 284-293.

MITTLER, G., F. BUTTER and M. MANN, 2009b A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome Res.

MIZUNO, S., T. CHIJIWA, T. OKAMURA, K. AKASHI, Y. FUKUMAKI *et al.*, 2001 Expression of DNA methyltransferases DNMT1, 3A, and 3B in normal hematopoiesis and in acute and chronic myelogenous leukemia. Blood **97:** 1172-1179.

MOHN, F., M. WEBER, M. REBHAN, T. C. ROLOFF, J. RICHTER *et al.*, 2008 Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular cell **30:** 755-766.

MORAN-CRUSIO, K., L. REAVIE, A. SHIH, O. ABDEL-WAHAB, D. NDIAYE-LOBRY *et al.*, 2011 Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer Cell **20:** 11-24.

MORGAN, H. D., W. DEAN, H. A. COKER, W. REIK and S. K. PETERSEN-MAHRT, 2004 Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. The Journal of biological chemistry **279:** 52353-52360.

MORGAN, H. D., F. SANTOS, K. GREEN, W. DEAN and W. REIK, 2005 Epigenetic reprogramming in mammals. Human molecular genetics **14 Spec No 1:** R47-58.

MULLER-TIDOW, C., K. KUGLER, S. DIEDERICHS, S. KLUMPEN, M. MOLLER *et al.*, 2001 Loss of expression of HDAC-recruiting methyl-CpG-binding domain proteins in human cancer. British journal of cancer **85:** 1168-1174.

NG, M. H., R. K. NG, C. T. KONG, D. Y. JIN and L. C. CHAN, 2010 Activation of Ras-dependent Elk-1 activity by MLL-AF4 family fusion oncoproteins. Experimental hematology **38:** 481-488.

O'BRIEN, S., 2008 New agents in the treatment of CLL. Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program**:** 457-464.

OAKELEY, E. J., 1999 DNA methylation analysis: a review of current methodologies. Pharmacology & therapeutics **84:** 389-400.

OGINO, S., K. NOSHO, G. J. KIRKNER, T. KAWASAKI, A. T. CHAN *et al.*, 2008 A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. Journal of the National Cancer Institute **100:** 1734-1738.

OHM, J. E., K. M. MCGARVEY, X. YU, L. CHENG, K. E. SCHUEBEL *et al.*, 2007 A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet **39:** 237-242.

OKA, M., N. RODIC, J. GRADDY, L. J. CHANG and N. TERADA, 2006 CpG sites preferentially methylated by Dnmt3a in vivo. The Journal of biological chemistry **281:** 9901-9908.

OKANO, M., D. W. BELL, D. A. HABER and E. LI, 1999 DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell **99:** 247-257.

OKANO, M., S. XIE and E. LI, 1998 Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nature genetics **19:** 219-220.

OOI, S. K., C. QIU, E. BERNSTEIN, K. LI, D. JIA *et al.*, 2007 DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature **448:** 714-717.

OSBORN, L., S. KUNKEL and G. J. NABEL, 1989 Tumor necrosis factor alpha and interleukin 1 stimulate the human immunodeficiency virus enhancer by activation of the nuclear factor kappa B. Proceedings of the National Academy of Sciences of the United States of America **86:** 2336-2340.

PAL, R., M. JANZ, D. L. GALSON, M. GRIES, S. LI *et al.*, 2009 C/EBPbeta regulates transcription factors critical for proliferation and survival of multiple myeloma cells. Blood **114:** 3890-3898.

PASTOR, W. A., U. J. PAPE, Y. HUANG, H. R. HENDERSON, R. LISTER *et al.*, 2011 Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature **473:** 394-397.

PATEL, K., J. DICKSON, S. DIN, K. MACLEOD, D. JODRELL *et al.*, 2010 Targeting of 5-aza-2'-deoxycytidine residues by chromatin-associated DNMT1 induces proteasomal degradation of the free enzyme. Nucleic acids research **38:** 4313-4324.

PAZ, M. F., M. F. FRAGA, S. AVILA, M. GUO, M. POLLAN *et al.*, 2003 A systematic profile of DNA methylation in human cancer cell lines. Cancer research **63:** 1114-1121.

PEKARSKY, Y., C. HALLAS and C. M. CROCE, 2003 Targeting mature T cell leukemia: new understanding of molecular pathways. American journal of pharmacogenomics : genomics-related research in drug development and clinical practice **3:** 31-36.

PELIZZOLA, M., Y. KOGA, A. E. URBAN, M. KRAUTHAMMER, S. WEISSMAN *et al.*, 2008 MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome Res **18:** 1652-1659.

PLASS, C., J. C. BYRD, A. RAVAL, S. M. TANNER and A. DE LA CHAPELLE, 2007 Molecular profiling of chronic lymphocytic leukaemia: genetics meets epigenetics to identify predisposing genes. Br J Haematol **139:** 744-752.

POLAK, P., and E. DOMANY, 2006 Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics **7:** 133.

POLLACK, Y., R. STEIN, A. RAZIN and H. CEDAR, 1980 Methylation of foreign DNA sequences in eukaryotic cells. Proceedings of the National Academy of Sciences of the United States of America **77:** 6463-6467.

POPP, C., W. DEAN, S. FENG, S. J. COKUS, S. ANDREWS *et al.*, 2010 Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature **463:** 1101-1105.

QUINLIVAN, E. P., and J. F. GREGORY, 3RD, 2008 DNA methylation determination by liquid chromatography-tandem mass spectrometry using novel biosynthetic [U-15N]deoxycytidine and [U-15N]methyldeoxycytidine internal standards. Nucleic acids research **36:** e119.

RAHMATPANAH, F. B., S. CARSTENS, S. I. HOOSHMAND, E. C. WELSH, O. SJAHPUTERA *et al.*, 2009 Large-scale analysis of DNA methylation in chronic lymphocytic leukemia. Epigenomics **1:** 39-61.

RAI, K., I. J. HUGGINS, S. R. JAMES, A. R. KARPF, D. A. JONES *et al.*, 2008 DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. Cell **135:** 1201-1212.

RAKYAN, V. K., T. A. DOWN, N. P. THORNE, P. FLICEK, E. KULESHA *et al.*, 2008 An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Res **18:** 1518-1529.

RAMSAHOYE, B. H., D. BINISZKIEWICZ, F. LYKO, V. CLARK, A. P. BIRD *et al.*, 2000 Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proceedings of the National Academy of Sciences of the United States of America **97:** 5237-5242.

RAUCH, T., and G. P. PFEIFER, 2005 Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. Lab Invest **85:** 1172-1180.

RAUCH, T., Z. WANG, X. ZHANG, X. ZHONG, X. WU *et al.*, 2007 Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. Proc Natl Acad Sci U S A **104:** 5527-5532.

RAUCH, T. A., and G. P. PFEIFER, 2010 DNA methylation profiling using the methylated-CpG island recovery assay (MIRA). Methods **52:** 213-217.

RAUCH, T. A., X. WU, X. ZHONG, A. D. RIGGS and G. P. PFEIFER, 2009 A human B cell methylome at 100-base pair resolution. Proc Natl Acad Sci U S A **106:** 671-678.

RAUCH, T. A., X. ZHONG, X. WU, M. WANG, K. H. KERNSTINE *et al.*, 2008 High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. Proc Natl Acad Sci U S A **105:** 252-257.

RAVAL, A., D. M. LUCAS, J. J. MATKOVIC, K. L. BENNETT, S. LIYANARACHCHI *et al.*, 2005 TWIST2 demonstrates differential methylation in immunoglobulin variable heavy chain mutated and unmutated chronic lymphocytic leukemia. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **23:** 3877-3885.

RAVAL, A., S. M. TANNER, J. C. BYRD, E. B. ANGERMAN, J. D. PERKO *et al.*, 2007 Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. Cell **129:** 879-890.

RAWSTRON, A. C., 2009 Monoclonal B-cell lymphocytosis. Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program**:** 430-439.

REINDERS, J., C. DELUCINGE VIVIER, G. THEILER, D. CHOLLET, P. DESCOMBES *et al.*, 2008 Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. Genome research **18:** 469-476.

ROBERTSON, K. D., S. AIT-SI-ALI, T. YOKOCHI, P. A. WADE, P. L. JONES *et al.*, 2000 DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. Nature genetics **25:** 338-342.

ROBERTSON, K. D., E. UZVOLGYI, G. LIANG, C. TALMADGE, J. SUMEGI *et al.*, 1999 The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. Nucleic acids research **27:** 2291-2298.

ROHDE, C., Y. ZHANG, R. REINHARDT and A. JELTSCH, 2010 BISMA--fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. BMC Bioinformatics **11:** 230.

ROLLINS, R. A., F. HAGHIGHI, J. R. EDWARDS, R. DAS, M. Q. ZHANG *et al.*, 2006 Large-scale structure of genomic methylation patterns. Genome Res **16:** 157-163.

ROSS, J. P., K. N. RAND and P. L. MOLLOY, 2010 Hypomethylation of repeated DNA sequences in cancer. Epigenomics **2:** 245-269.

ROUPRET, M., V. HUPERTAN, D. R. YATES, J. W. CATTO, I. REHMAN *et al.*, 2007 Molecular detection of localized prostate cancer using quantitative methylation-specific PCR on urinary cells obtained following prostate massage. Clinical cancer research : an official journal of the American Association for Cancer Research **13:** 1720-1725.

RUSH, L. J., Z. DAI, D. J. SMIRAGLIA, X. GAO, F. A. WRIGHT *et al.*, 2001 Novel methylation targets in de novo acute myeloid leukemia with prevalence of chromosome 11 loci. Blood **97:** 3226-3233.

RUSH, L. J., A. RAVAL, P. FUNCHAIN, A. J. JOHNSON, L. SMITH *et al.*, 2004 Epigenetic profiling in chronic lymphocytic leukemia reveals novel methylation targets. Cancer research **64:** 2424-2433.

SAITO, K., K. KAWAKAMI, I. MATSUMOTO, M. ODA, G. WATANABE *et al.*, 2010 Long interspersed nuclear element 1 hypomethylation is a marker of poor prognosis in stage IA non-small cell lung cancer. Clinical cancer research : an official journal of the American Association for Cancer Research **16:** 2418-2426.

SAITO, M., and F. ISHIKAWA, 2002 The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. The Journal of biological chemistry **277:** 35434-35439.

SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular Cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

SANFORD, J., L. FORRESTER, V. CHAPMAN, A. CHANDLEY and N. HASTIE, 1984 Methylation patterns of repetitive DNA sequences in germ cells of Mus musculus. Nucleic acids research **12:** 2823-2836.

SASAI, N., M. NAKAO and P. A. DEFOSSEZ, 2010 Sequence-specific recognition of methylated DNA by human zinc-finger proteins. Nucleic acids research **38:** 5015-5022.

SAWADA, M., Y. KANAI, E. ARAI, S. USHIJIMA, H. OJIMA *et al.*, 2007 Increased expression of DNA methyltransferase 1 (DNMT1) protein in uterine cervix squamous cell carcinoma and its precursor lesion. Cancer letters **251:** 211-219.

SAXONOV, S., P. BERG and D. L. BRUTLAG, 2006 A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A **103:** 1412-1417.

SCHLESINGER, Y., R. STRAUSSMAN, I. KESHET, S. FARKASH, M. HECHT *et al.*, 2007 Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet **39:** 232-236.

SCHMITZ, K. M., N. SCHMITT, U. HOFFMANN-ROHRER, A. SCHAFER, I. GRUMMT *et al.*, 2009 TAF12 recruits Gadd45a and the nucleotide excision repair complex to the promoter of rRNA genes leading to active DNA demethylation. Molecular cell **33:** 344-353.

SCHROEDER, H. W., JR., and G. DIGHIERO, 1994 The pathogenesis of chronic lymphocytic leukemia: analysis of the antibody repertoire. Immunology today **15:** 288-294.

SCHUH, K., A. AVOTS, H. P. TONY, E. SERFLING and C. KNEITZ, 1996 Nuclear NF-ATp is a hallmark of unstimulated B cells from B-CLL patients. Leukemia & lymphoma **23:** 583-592.

SEELIGER, B., S. WILOP, R. OSIEKA, O. GALM and E. JOST, 2009 CpG island methylation patterns in chronic lymphocytic leukemia. Leuk Lymphoma **50:** 419-426.

SHAH, M. Y., and J. D. LICHT, 2011 DNMT3A mutations in acute myeloid leukemia. Nature genetics **43:** 289-290.

SHAO, C., M. LACEY, L. DUBEAU and M. EHRLICH, 2009 Hemimethylation footprints of DNA demethylation in cancer. Epigenetics **4:** 165-175.

SHARMA, S., D. D. DE CARVALHO, S. JEONG, P. A. JONES and G. LIANG, 2011 Nucleosomes containing methylated DNA stabilize DNA methyltransferases 3A/3B and ensure faithful epigenetic inheritance. PLoS genetics **7:** e1001286.

SHEARSTONE, J. R., R. POP, C. BOCK, P. BOYLE, A. MEISSNER *et al.*, 2011 Global DNA demethylation during mouse erythropoiesis in vivo. Science **334:** 799-802.

SHEN, L., Y. KONDO, Y. GUO, J. ZHANG, L. ZHANG *et al.*, 2007 Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS Genet **3:** 2023-2036.

SHIOTA, K., 2004 DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. Cytogenet Genome Res **105:** 325-334.

SILKE, J., K. I. ROTHER, O. GEORGIEV, W. SCHAFFNER and K. MATSUO, 1995 Complex demethylation patterns at Sp1 binding sites in F9 embryonal carcinoma cells. FEBS letters **370:** 170-174.

SINGER, J., J. ROBERTS-EMS, F. W. LUTHARDT and A. D. RIGGS, 1979 Methylation of DNA in mouse early embryos, teratocarcinoma cells and adult tissues of mouse and rabbit. Nucleic acids research **7:** 2369-2385.

SMALLWOOD, A., P. O. ESTEVE, S. PRADHAN and M. CAREY, 2007 Functional cooperation between HP1 and DNMT1 mediates gene silencing. Genes & development **21:** 1169-1178.

SMIRAGLIA, D. J., R. KAZHIYUR-MANNAR, C. C. OAKES, Y. Z. WU, P. LIANG *et al.*, 2007 Restriction landmark genomic scanning (RLGS) spot identification by second generation virtual RLGS in multiple genomes with multiple enzyme combinations. BMC Genomics **8:** 446.

SMIRAGLIA, D. J., L. J. RUSH, M. C. FRUHWALD, Z. DAI, W. A. HELD *et al.*, 2001 Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. Human molecular genetics **10:** 1413-1419.

SOINI, T., H. HAVERI, J. M. ELO, M. KAUPPINEN, A. KYRONLAHTI *et al.*, 2011 Transcription Factor GATA-4 Is Abundantly Expressed In Childhood But Not In Adult Liver Tumors. Journal of pediatric gastroenterology and nutrition.

SONG, F., J. F. SMITH, M. T. KIMURA, A. D. MORROW, T. MATSUYAMA *et al.*, 2005 Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc Natl Acad Sci U S A **102:** 3336-3341.

SONG, J., O. RECHKOBLIT, T. H. BESTOR and D. J. PATEL, 2011 Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science **331:** 1036-1040.

SONG, J. H., H. J. KIM, C. H. LEE, S. J. KIM, S. Y. HWANG *et al.*, 2006 Identification of gene expression signatures for molecular classification in human leukemia cells. International journal of oncology **29:** 57-64.

SPROUL, D., C. NESTOR, J. CULLEY, J. H. DICKSON, J. M. DIXON *et al.*, 2011 Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. Proceedings of the National Academy of Sciences of the United States of America **108:** 4364-4369.

STRUNNIKOVA, M., U. SCHAGDARSURENGIN, A. KEHLEN, J. C. GARBE, M. R. STAMPFER *et al.*, 2005 Chromatin inactivation precedes de novo DNA methylation during

the progressive epigenetic silencing of the RASSF1A promoter. Molecular and cellular biology **25:** 3923-3933.

SUETAKE, I., F. SHINOZAKI, J. MIYAGAWA, H. TAKESHIMA and S. TAJIMA, 2004 DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. The Journal of biological chemistry **279:** 27816-27823.

SUMMERER, D., 2010 High-throughput DNA sequencing beyond the four-letter code: epigenetic modifications revealed by single-molecule bypass kinetics. Chembiochem : a European journal of chemical biology **11:** 2499-2501.

SUMNER, A. T., A. R. MITCHELL and P. M. ELLIS, 1998 A FISH study of chromosome fusion in the ICF syndrome: involvement of paracentric heterochromatin but not of the centromeres themselves. Journal of medical genetics **35:** 833-835.

SUN, Y. V., S. T. TURNER, J. A. SMITH, P. I. HAMMOND, A. LAZARUS *et al.*, 2010 Comparison of the DNA methylation profiles of human peripheral blood cells and transformed B-lymphocytes. Hum Genet **127:** 651-658.

SUTER, C. M., D. I. MARTIN and R. L. WARD, 2004 Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue. Int J Colorectal Dis **19:** 95-101.

SUZUKI, M. M., and A. BIRD, 2008 DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet **9:** 465-476.

SZYF, M., V. BOZOVIC and G. TANIGAWA, 1991 Growth regulation of mouse DNA methyltransferase gene expression. The Journal of biological chemistry **266:** 10027-10030.

TACHIBANA, M., K. SUGIMOTO, M. NOZAKI, J. UEDA, T. OHTA *et al.*, 2002 G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. Genes & development **16:** 1779-1791.

TAHILIANI, M., K. P. KOH, Y. SHEN, W. A. PASTOR, H. BANDUKWALA *et al.*, 2009 Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science **324:** 930-935.

TAMURA, T., Y. ARAKI, S. YAMAOKA, K. INAGAKI and H. TANAKA, 1997 Determination of methylation specificity of sequence-specific DNA methyltransferases using matrix assisted laser desorption/ionization time-of-flight mass spectrometry. Nucleic acids research **25:** 4162-4164.

TEFFERI, A., K. H. LIM, O. ABDEL-WAHAB, T. L. LASHO, J. PATEL *et al.*, 2009 Detection of mutant TET2 in myeloid malignancies other than myeloproliferative neoplasms: CMML, MDS, MDS/MPN and AML. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K **23:** 1343-1345.

THOMPSON, R. F., G. ATZMON, C. GHEORGHE, H. Q. LIANG, C. LOWES *et al.*, 2010 Tissue-specific dysregulation of DNA methylation in aging. Aging Cell **9:** 506-518.

TIWARI, V. K., K. M. MCGARVEY, J. D. LICCHESI, J. E. OHM, J. G. HERMAN *et al.*, 2008 PcG proteins, DNA methylation, and gene repression by chromatin looping. PLoS Biol **6:** 2911-2927.

TONG, W. G., W. G. WIERDA, E. LIN, S. Q. KUANG, B. N. BEKELE *et al.*, 2010 Genome-wide DNA methylation profiling of chronic lymphocytic leukemia allows identification of epigenetically repressed molecular pathways with

clinical impact. Epigenetics : official journal of the DNA Methylation Society **5:** 499-508.

TUFARELLI, C., J. A. STANLEY, D. GARRICK, J. A. SHARPE, H. AYYUB *et al.*, 2003 Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat Genet **34:** 157-165.

VAKOC, C. R., S. A. MANDAT, B. A. OLENCHOCK and G. A. BLOBEL, 2005 Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. Molecular cell **19:** 381-391.

VALINLUCK, V., and L. C. SOWERS, 2007a Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. Cancer research **67:** 946-950.

VALINLUCK, V., and L. C. SOWERS, 2007b Inflammation-mediated cytosine damage: a mechanistic link between inflammation and the epigenetic alterations in human cancers. Cancer research **67:** 5583-5586.

VAN DEN BROECK, A., E. BRAMBILLA, D. MORO-SIBILOT, S. LANTUEJOUL, C. BRAMBILLA *et al.*, 2008 Loss of histone H4K20 trimethylation occurs in preneoplasia and influences prognosis of non-small cell lung cancer. Clinical cancer research : an official journal of the American Association for Cancer Research **14:** 7237-7245.

VAN NIMWEGEN, E., 2007 Finding regulatory elements and regulatory motifs: a general probabilistic framework. BMC Bioinformatics **8 Suppl 6:** S4.

VIRE, E., C. BRENNER, R. DEPLUS, L. BLANCHON, M. FRAGA *et al.*, 2006 The Polycomb group protein EZH2 directly controls DNA methylation. Nature **439:** 871-874.

WAHLFORS, J., H. HILTUNEN, K. HEINONEN, E. HAMALAINEN, L. ALHONEN *et al.*, 1992 Genomic hypomethylation in human chronic lymphocytic leukemia. Blood **80:** 2074-2080.

WALSH, C. P., J. R. CHAILLET and T. H. BESTOR, 1998 Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet **20:** 116-117.

WANG, J., K. R. COOMBES, W. E. HIGHSMITH, M. J. KEATING and L. V. ABRUZZO, 2004 Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. Bioinformatics **20:** 3166-3178.

WANG, J., S. HEVI, J. K. KURASH, H. LEI, F. GAY *et al.*, 2009 The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation. Nature genetics **41:** 125-129.

WARNECKE, P. M., C. STIRZAKER, J. R. MELKI, D. S. MILLAR, C. L. PAUL *et al.*, 1997 Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic acids research **25:** 4422-4426.

WARNECKE, P. M., C. STIRZAKER, J. SONG, C. GRUNAU, J. R. MELKI *et al.*, 2002 Identification and resolution of artifacts in bisulfite sequencing. Methods **27:** 101-107.

WEBER, M., J. J. DAVIES, D. WITTIG, E. J. OAKELEY, M. HAASE *et al.*, 2005 Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet **37:** 853-862.

WEBER, M., I. HELLMANN, M. B. STADLER, L. RAMOS, S. PAABO *et al.*, 2007 Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet **39:** 457-466.

WEBER, M., and D. SCHUBELER, 2007 Genomic patterns of DNA methylation: targets and function of an epigenetic mark. Curr Opin Cell Biol **19:** 273-280.

WEISENBERGER, D. J., M. CAMPAN, T. I. LONG, M. KIM, C. WOODS *et al.*, 2005 Analysis of repetitive element DNA methylation by MethyLight. Nucleic acids research **33:** 6823-6836.

WEISS, A., and H. CEDAR, 1997 The role of DNA demethylation during development. Genes Cells **2:** 481-486.

WIDSCHWENDTER, M., H. FIEGL, D. EGLE, E. MUELLER-HOLZNER, G. SPIZZO *et al.*, 2007 Epigenetic stem cell signature in cancer. Nature genetics **39:** 157-158.

WIENHOLZ, B. L., M. S. KARETA, A. H. MOAREFI, C. A. GORDON, P. A. GINNO *et al.*, 2010 DNMT3L modulates significant and distinct flanking sequence preference for DNA methylation by DNMT3A and DNMT3B in vivo. PLoS genetics **6**.

WIGLER, M., D. LEVY and M. PERUCHO, 1981 The somatic replication of DNA methylation. Cell **24:** 33-40.

WILD, L., and J. M. FLANAGAN, 2010 Genome-wide hypomethylation in cancer may be a passive consequence of transformation. Biochim Biophys Acta **1806:** 50-57.

WILKIE, A. O., H. C. ZEITLIN, R. H. LINDENBAUM, V. J. BUCKLE, N. FISCHEL-GHODSIAN *et al.*, 1990 Clinical features and molecular analysis of the alpha thalassemia/mental retardation syndromes. II. Cases without detectable abnormality of the alpha globin complex. Am J Hum Genet **46:** 1127-1140.

WOODFINE, K., J. E. HUDDLESTON and A. MURRELL, 2011 Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. Epigenetics & chromatin **4:** 1.

WOSSIDLO, M., T. NAKAMURA, K. LEPIKHOV, C. J. MARQUES, V. ZAKHARTCHENKO *et al.*, 2011 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nature communications **2:** 241.

WYATT, G. R., and S. S. COHEN, 1952 A new pyrimidine base from bacteriophage nucleic acids. Nature **170:** 1072-1073.

WYRZYKOWSKA, P., K. STALINSKA, M. WAWRO, J. KOCHAN and A. KASZA, 2010 Epidermal growth factor regulates PAI-1 expression via activation of the transcription factor Elk-1. Biochimica et biophysica acta **1799:** 616-621.

XI, Y., and W. LI, 2009 BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics **10:** 232.

XIE, H., M. WANG, A. DE ANDRADE, F. BONALDO MDE, V. GALAT *et al.*, 2011 Genome-wide quantitative assessment of variation in DNA methylation patterns. Nucleic acids research **39:** 4099-4108.

XU, G. L., T. H. BESTOR, D. BOURC'HIS, C. L. HSIEH, N. TOMMERUP *et al.*, 1999 Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature **402:** 187-191.

YAMASHITA, Y., J. YUAN, I. SUETAKE, H. SUZUKI, Y. ISHIKAWA *et al.*, 2010 Array-based genomic resequencing of human leukemia. Oncogene **29:** 3723-3731.

YAN, X. J., J. XU, Z. H. GU, C. M. PAN, G. LU *et al.*, 2011 Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nature genetics **43:** 309-315.

YASUI, D. H., S. PEDDADA, M. C. BIEDA, R. O. VALLERO, A. HOGART *et al.*, 2007 Integrated epigenomic analyses of neuronal MeCP2 reveal a role for long-range interaction with active genes. Proceedings of the National Academy of Sciences of the United States of America **104:** 19416-19421.

YEGNASUBRAMANIAN, S., M. C. HAFFNER, Y. ZHANG, B. GUREL, T. C. CORNISH *et al.*, 2008 DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity. Cancer research **68:** 8954-8967.

YILMAZ, M., D. MAASS, N. TIWARI, L. WALDMEIER, P. SCHMIDT *et al.*, 2011 Transcription factor Dlx2 protects from TGFbeta-induced cell-cycle arrest and apoptosis. The EMBO journal **30:** 4489-4499.

YODER, J. A., C. P. WALSH and T. H. BESTOR, 1997 Cytosine methylation and the ecology of intragenomic parasites. Trends in genetics : TIG **13:** 335-340.

YOUNG, J. I., E. P. HONG, J. C. CASTLE, J. CRESPO-BARRETO, A. B. BOWMAN *et al.*, 2005 Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. Proceedings of the National Academy of Sciences of the United States of America **102:** 17551-17558.

YU, M. K., 2006 Epigenetics and chronic lymphocytic leukemia. Am J Hematol **81:** 864-869.

YU, M. K., H. BERGONIA, A. SZABO and J. D. PHILLIPS, 2007 Progressive disease in chronic lymphocytic leukemia is correlated with the DNA methylation index. Leukemia research **31:** 773-777.

ZEMACH, A., I. E. MCDANIEL, P. SILVA and D. ZILBERMAN, Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science **328:** 916-919.

ZEMOJTEL, T., S. M. KIELBASA, P. F. ARNDT, S. BEHRENS, G. BOURQUE *et al.*, 2011 CpG deamination creates transcription factor binding sites with high efficiency. Genome biology and evolution.

ZHANG, R. P., J. Z. SHAO and L. X. XIANG, 2011 GADD45A Protein Plays an Essential Role in Active DNA Demethylation during Terminal Osteogenic Differentiation of Adipose-derived Mesenchymal Stem Cells. The Journal of biological chemistry **286:** 41083-41094.

ZHAO, Q., G. RANK, Y. T. TAN, H. LI, R. L. MORITZ *et al.*, 2009 PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. Nature structural & molecular biology **16:** 304-311.

ZHENG, Z., S. VENKATAPATHY, G. RAO and C. A. HARRINGTON, 2002 Expression profiling of B cell chronic lymphocytic leukemia suggests deficient CD1-mediated immunity, polarized cytokine response, altered adhesion and increased intracellular protein transport and processing of leukemic cells. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K **16:** 2429-2437.

ZHU, B., Y. ZHENG, H. ANGLIKER, S. SCHWARZ, S. THIRY *et al.*, 2000a 5-Methylcytosine DNA glycosylase activity is also present in the human MBD4 (G/T mismatch glycosylase) and in a related avian sequence. Nucleic acids research **28:** 4157-4165.

ZHU, D., J. FANG, Y. LI and J. ZHANG, 2009 Mbd3, a component of NuRD/Mi-2 complex, helps maintain pluripotency of mouse embryonic stem cells by repressing trophectoderm differentiation. PLoS ONE **4:** e7684.

ZHU, Z., S. LING, Q. H. YANG and L. LI, 2000b The difference in the carboxy-terminal sequence is responsible for the difference in the activity of chicken and rat liver fructose-2,6-bisphosphatase. Biological chemistry **381:** 1195-1202.

ZILBERMAN, D., D. COLEMAN-DERR, T. BALLINGER and S. HENIKOFF, 2008 Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature **456:** 125-129.

ZILBERMAN, D., M. GEHRING, R. K. TRAN, T. BALLINGER and S. HENIKOFF, 2007 Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nature genetics **39:** 61-69.